AN INVESTIGATION OF THE FIT OF THE RASCH MEASUREMENT MODEL TO DATA FROM THE MEDICAL COLLEGE ADMISSION TEST

DISSERTATION FOR THE DEGREE OF PH.D.

MICHIGAN STATE UNIVERSITY

GILLES CORMIER

1977

This is to certify that the

thesis entitled

AN INVESTIGATION OF THE FIT OF THE RASCH MEASUREMENT MODEL TO DATA FROM THE MEDICAL COLLEGE ADMISSION TEST

presented by

GILLES CORMIER

has been accepted towards fulfillment
of the requirements for

PH. D degree in EDUCATIONAL PSYCHOLOGY

William H. Schmidt

Major professor

Date MARCH 4, 1977

139 88

ABSTRACT

AN INVESTIGATION OF THE FIT OF THE RASCH MEASUREMENT MODEL TO DATA FROM THE MEDICAL COLLEGE ADMISSION TEST

By

Gilles Cormier

The main purpose of this study is to examine a measurement model, to explore the range of situations to which it can be applied, and to establish its value in the improvement of the decision-making process in education. Considering the inadequacies of the classical theory of testing in many situations, the measurement model proposed here and developed by Georg Rasch comprises some interesting features. It allows for sample-free test calibration and sample-free person measurement. In other words, contrary to the classical linear model, the Rasch model provides item and ability parameters that remain invariant as the item analysis group changes. Such a model relies on some assumptions. How robust is the model to violations of those assumptions? What characteristics must a test have so that there is fit between model and data? What does the very notion of fit mean? How can it be established? These are some of the questions this study examines in the light of other already published papers on the same issue. A major objective of this study is thus to add new data to the available body of knowledge concerning the simple logistic model developed by Rasch so that the goal of achieving meaningful measurement through invariant scaling and objectivity becomes possible.

The basic requirements needed for conducting the investigations presented here were found in the data from the Medical College Admission

Test (MCAT). The analyses were performed on three of the four MCAT subtests administered in May of 1972 (18,075 subjects): Verbal Ability Subtest (75 items), Quantitative Ability Subtest (50 items), and Science Subtest (86 items).

Two series of analyses were conducted. In the first series, an attempt is made to find out whether or not there is fit between the three MCAT subtests and the simple logistic model using a chi-square test of fit applied to the overall test and to each item. The most likely hypotheses of misfit, that is, item discrimination, guessing and speed are then examined. In the second series, the effects of misfit on test calibration and person measurement are assessed. To achieve this, the sample of examinees was divided into various subgroups, the test was calibrated on each subgroup and the magnitude of the differences between easiness and ability estimates was evaluated.

The major findings of this study concern the MCAT itself and the notion of model-data fit. As for the MCAT, the results illustrate that estimates of ability are free from sample considerations under the simple logistic model, and thus, that the Rasch model fits the MCAT data, that is, applies to its three aptitude subtests and is not influenced by three of its population's characteristics: intellectual ability, socio-economic status, and race. As for the concept of model-data fit, an important relationship is established in this study between criteria of fit at the item level and indicators of fit at the test level.

AN INVESTIGATION OF THE FIT OF THE RASCH MEASUREMENT MODEL TO DATA FROM THE MEDICAL COLLEGE ADMISSION TEST

By

Gilles Cormier

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Personal Services
and Educational Psychology
College of Education

1977

ACKNOWLEDGMENTS

My sincere thanks are due to the members of my Committee, Professors Lee Shulman, Maryellen McSweeny, and Howard Teitelbaum for their encouragement and help during the preparation of this dissertation.

Invaluable assistance was provided and helpful suggestions were made by Professor William Schmidt, Chairman of my Committee, throughout the entire course of this work. Without his constant support during the many years it took to complete this dissertation, it is doubtful that it could have been brought to an end.

I wish to thank the Medical Research Council of Canada for the Fellowship granted to me during the three years of my doctoral training. Financial support was also provided by the Association of American Medical Colleges more specifically for this dissertation. I am grateful to my dear friends James Erdmann and Ayres D'Costa for making this arrangement possible.

Thanks are also due to my wife and children for their understanding and patience throughout the various phases of this project.

Finally, I owe a special debt of gratitude to Jocelyne Lessard for her tireless assistance in the typing of the many drafts required before this final presentation.

TABLE OF CONTENTS

TABLE OF CONTENTS - Continued

LIST OF TABLES

LIST OF TABLES - Continued

LIST OF TABLES - Continued

LIST OF TABLES - Continued

LIST OF TABLES - Continued

CHAPTER I

STATEMENT OF THE PROBLEM

The main purpose of this study is to examine a measurement model, to explore the range of situations to which it can be applied, and to establish its value in the improvement of the decision-making process in medical education. The viewpoint emphasized here is that of a practitioner. The objective of this set of investigations is thus to help fill the gap between theory and practice.

Some major problems have been known to exist for a long time in ability measurement. With the exception of a few well standardized tests, most of the actual testing in education relies heavily on observations made on non-calibrated instruments. To illustrate how unsatisfactory such a practice could be, consider the degree of confidence a physician would have in recommending that a group of patients follow a diet for the purpose of losing weight when the basis for the recommendation is a series of observations made on a non-calibrated scale, that is, a scale that has no absolute zero and unequal units of unknown length.

Students are certified to practice medicine on the basis of how well they do on certain tests. Since there are no known valid criteria of performance in practice to which scores on these tests can be related, decisions are based on group-centered statistics. Expressed rather bluntly this amounts to saying that the chances a student has of becoming a physician are a function of the yearly group to which he

happens to belong and the particular set of tests he happens to take. A tight selection process and past experience with students act as safeguards to prevent major disasters and to ensure relative public safety but it is this thesis contention that the method is basically wrong.

There is a definite trend in medical education to develop pools of items in many areas of achievement. It is expected that with a very large number of items and an adequate sampling procedure several tests can be constructed which can then be administered to different groups of students and yield comparable results. However the methodology presently used to put the universe of items on a common scale lacks a sound theoretical base and scientific rigor. Consequently, parallel sets of items drawn from these pools and administered to the same group of subjects rarely yield comparable estimates of ability when the expected error of measurement is taken into account. Such a state of affairs is unsatisfactory.

Another important trend in medical education has to do with national and even international assessment of medical competency. Such a goal cannot be achieved within the current framework of ability testing. It follows logically that if different instruments yield different estimates of ability for the same subjects meaningful comparative evaluation can hardly be hoped for.

These problems stem from the inadequacies of the paradigm presently used in measurement. In the classical theory of testing the two stages of measurement, test calibration and ability estimation, are confounded. The tests are calibrated on the basis of some samples of subjects. Two item indices are then derived which are sample bound:

item discrimination and item difficulty. Item discrimination is a measure of the quality of the item, that is, how well the item differentiates between subjects of different ability. Item difficulty is simply the proportion of people that answer an item correctly. Because of their dependence on the sample used to derive them, these indices are not invariant. A new sample of subjects whose range of ability is different will provide different values for these two indices. Furthermore, the total test score is used as an estimate of each subject's ability. Such an unreliable measure which is at best on an ordinal scale does not allow for meaningful comparisons between subjects. The deficiencies of the classical model have long been acknowledged. Gulliksen who contributed to shaping the classical theory of testing some twenty-five years ago expressed his concern clearly.

> "An important contribution to item analysis theory would be the discovery of item parameters that remained relatively stable as the item analysis group changed." (Gulliksen, 1950, p. 392).

The model proposed in this study represents such a contribution. It allows for objectivity in measurement. This particular kind of objectivity however, requires two conditions. First, the calibration of the test must be independent of the particular group of subjects used for the calibration. Second, the measurement of subjects must be independent of the particular test used for the measurement.

> "If mental measurement has to have any meaning in a scientific context a different approach which allows for objective measurement must be used." (Panchapakesan, 1969).

Current practices do not recognize the double nature of the measurement process and attempts are made to measure a given trait with different instruments possessing no common scale. It has already

been stated that this was an unsatisfactory state of affairs. If a choice were possible, the educational practitioner would be expected to adopt, among competing models, the one that best approximates the ideal situation of objectivity, that is, a model that would allow for sample-free test calibration and sample-free person measurement.

Fortunately such a choice is possible. In the mid-50's, Georg Rasch, a mathematician at the Danish Institute for Educational Research, developed the theory that led to the formulation of the model studied here and referred to as the simple logistic model. The form of the model, its assumptions, its estimation procedure, its test of fit, and how it allows for objectivity in measurement are subjects dealt with in Chapter II. The simple logistic model provides a way of scaling tests so that instead of using the sample dependent test score as a measure of ability, one can use an invariant measure, invariant over tests of the same ability.

But in spite of the fact that such a model exists, it has not come to be used generally. The classical test model entails so many deficiencies that one would expect that if there were a procedure based on a sound theory capable of solving measurement problems which could not be solved as efficiently otherwise such a procedure would be quickly adopted. Wright (1968) suggested an answer.

> "Perhaps too few recognize the importance of objectivity in mental measurement. Perhaps, too, many despair that it can ever be achieved, or fear it will be too difficult to do. What we need is some evidence that objective measurements of mental ability can really be made." (Wright, 1968).

We feel that it is unreasonable to expect major changes in present practices until Rasch's model is proven to be a valuable

alternative. For this to happen, many studies need be done to span the whole range of situations to which the model can be applied. An overview of the studies of the fit of this model to various sets of data, real and simulated, which have been published so far is presented in Chapter III.

A major objective of this study is to add new data to the available body of knowledge concerning the simple logistic model so that the goal of achieving meaningful measurement through invariant scaling and objectivity becomes possible. Panchapakesan (1969) has successfully demonstrated the validity of simulation for the investigation of the robustness of the simple logistic model. However, she suggested that her technique which consisted of studying the effect of one departure from the model at a time should be checked with real data.

> "Real data will deviate from the model in several ways. Whether the various departures from the model can be disentangled is a question not answered by this study. All that can be said is that if different violations of the assumptions of the model do not confound each other it should be possible to apply the model successfully to real data." (Panchapakesan, 1969, p. 183).

This study follows up on Panchapakesan's work. Its main purpose is to suggest new ways to systematically explore the sources and the causes of misfit of the simple logistic model to real data. The data to test the fit of the model will be the three subtests of the Medical College Admission Test (MCAT) which comprises a set of quality aptitude tests used by medical schools in selecting their students. Chapter IV provides a description of the test, the sample, and the procedures used for this study.

The possible sources of misfit are numerous. We first eliminated the most obvious one, mis-scoring, by a careful analysis of the

suggested right answers. We then proceeded to weigh the influence of other factors according to a step-wise procedure recommended by Wright (1969). For the most challenged assumptions of the model concerning item discrimination, guessing and speed, we followed the plan adopted by Panchapakesan(1969) for her investigations with simulated data. The results of our analyses of each possible source of misfit of the model are presented in Chapter V.

The demonstration of the robustness of the model to violations of its assumptions is the subject considered in Chapter VI. That chapter is an attempt to assess the effects of some departures from the specifications of the simple logistic model on ability estimation and test calibration. No answer can be given to a question of whether the simple logistic model is valid or not. In measurement we know that a test is only valid for some specific purposes. The same rule applies to a model. A model is only valid for some specific situations. For what kinds of tests and what characteristics of a population of examinees is the simple logistic model valid? This is the crucial question this study addresses itself to. Because of the many features of the MCAT and the sizeable number of people who have taken it, a wide variety of interesting situations can be examined. However, the cost involved imposes some constraints.

Background

The inadequacies of the classical model lie within three major sources. First, the structure of the model itself, that is, the linear relationship postulated between an observed score and a true

score, cannot be properly tested since it is general enough to satisfy most data. A stronger model is needed to transform an observation into a measurement. Second, the measurement is not invariant since it depends upon a particular set of observations thus making any generalization over tests and samples totally unjustifiable. The two item statistics derived from the classical model are a function of the particular sample tested. The item easiness is a function of the average ability of the calibrating sample and the item discrimination depends upon the spread in the ability of the standardizing sample. Also the concept of reliability which is used as a measure of the worth of a test heavily relies on the variance in the scores of the standardizing sample. Third, the scale of measurement is so arbitrary that it is not amenable to easy interpretation. The metric used for the measure of a person's ability is simply the sum of item scores when each response is scored 0 or 1. The disadvantage of such a metric is that the score is largely a function of the particular test administered.

The earliest attempts to provide for some kind of invariant scaling were made by Thorndike (1926) and Thurstone (1925). They both tried to find a way of putting item estimates obtained from tests administered to samples of differing ability on a common scale. However, both approaches assumed that the ability distribution of each group tested was normal thus introducing a potential bias in the scale values obtained. Efforts to improve the classical linear model have failed to yield meaningful measurement. More powerful models were then sought. Lord's strong true score theory illustrates such a new trend (Lord, 1965) and so does a different conceptualization of the response of a subject to a stimulus as in quantal response models

(Lord and Novick, 1968). In these models, an attempt is made to express the relation between the observed response and the underlying ability which caused the response.

Any kind of deterministic model would be inappropriate in mental test theory since it would not produce an accurate description of the situation being observed. The uncertainty in the relationship between the parameters and the observations must be taken into account. This uncertainty could easily be reflected into a probabilistic statement. That is why a new class of models emphasizing such a notion was derived and referred to as probabilistic or stochastic. The two most popular models of that category are the normal model and the logistic model.

Stochastic response models have been applied to problems in other areas before being developed for mental testing. The normal ogive and the logistic models have been used to fit dosage response curves in bio-assay in order to assess the potency of vitamins, hormones, toxicants, and other drugs of all types by investigating the proportion of animals that succumb to a given dosage of a drug. Extensive analyses of dosage mortality curves have been carried out by Bliss (1935), Gaddum (1933), Finney (1952), Garwood (1941), and Berkson (1953). These studies have been reported by Panchapakesan (1969).

A more recent application to ROC (Receiving Operator Characteristic) curve-fitting has been reported by Grey and Morgan (1972) in the area of signal-detection. In these situations, a subject is presented with a stimulus z and is required to respond Yes (signal present) or No (signal not present). The ROC is a plot of P (Hit), the probability of striking right, against P (False alarm), the probability of making a

mistake, and is obtained by varying the stimulus z. P (Hit) is assumed to be equal to P(Yes|S), the conditional probability of saying Yes when the signal is present, and P (False alarm) is assumed to be equal to P(Yes|N), the conditional probability of saying Yes when only noise is present. The parameters of the ROC curve thus describe the subject's performance. The function relating the parameters to the observations has been found to be usually normal and sometimes logistic. Such a conceptualization might very well prove useful in the assessment of certain perceptual skills in a wide variety of situations in medical education (the interpretation of X-Ray films for instance).

The major difference between applications of these models to bio-assay and signal-detection on the one hand and to ability testing on the other hand lies in the fact that in the latter, ability is unknown and must be estimated, whereas in the former, either the dosage of a drug or the intensity of a stimulus can be manipulated. Since Chapter II deals with the development of such models in the area of ability measurement, we shall only point out, at this moment, that our choice for the logistic model over the normal model is based, partially at least, on a practical aspect of economy in time and labor. Baker (1961) made an empirical comparison of item parameters estimated on the logistic and normal models using the maximum likelihood method of estimation and found no significant difference between the two methods as expressed by the goodness of fit test. But the time required for the analysis with the logistic model was one-third that required for the analysis with the normal model.

In principle, the simple logistic model would be adequate only for tests made up of free-response items. However, the fact that the

MCAT is composed entirely of multiple choice items is not too disturbing given the evidence reported in the literature. For instance, Ross (1966) computed expected scores on the basis of the logistic model, assuming an underlying normal ability distribution, and obtained a reasonably good fit between the observed and the expected score distribution for two multiple choice tests and four free-response tests. We shall however examine this issue more extensively in Chapter III where we report on the most important studies of fit of the simple logistic model to different sets of data that were published in recent years.

CHAPTER II

THE SIMPLE LOGISTIC MODEL

Derivation of the model

The derivation of the simple logistic model can be looked at from two different perspectives which are going to be considered here successively. We shall first examine the train of thought followed by its author, and then, analyze the general class of models to which it belongs.

The simple logistic model is but one of the large number of models developed by the Danish mathematician Georg Rasch. From his involvement in the evaluation of a reading project for elementary school children in Denmark, Rasch's interest in mental measurement was triggered. Concerned with the lack of objectivity inherent in the classical linear test model, his efforts were aimed at finding more rigorous ways of achieving meaningful measurements. His contribution to what Wright (1968) calls the development of a science of measurement has long been ignored. The original research was published in a monograph now out of print (Rasch, 1960). A condensed treatment of the three most important Rasch models, two being Poisson process models (one for misreadings and the other for reading speed) and the third one being an item analysis model, is presented in Lord and Novick (1968, Chapter 21). Rasch's philosophy of measurement can be summarized by this quotation.

"A major characteristic of the Rasch models is that each is a

> two-parameter model with one parameter identified with the ability of the person and the second parameter identified with the difficulty of the measurement. Furthermore, the ability and difficulty parameters are separable in the Rasch models, so that they may be estimated independently... Finally, certain parameter-free distributions are available that make an investigation of the accuracy of the model possible." (Lord and Novick, 1968, pp. 480-81).

Of all the investigations of the empirical validity of these models reported by Rasch, it appears that the least favorable evidence is for the item analysis model. However, because of the desirable properties of such models, efforts should be encouraged to explore their practical uses.

> "It is clear that more detailed theoretical and more extensive empirical examinations of each of the Rasch models would be very useful, since these models, when their conditions are satisfied, provide a mode of analysis of great simplicity and power". (Lord and Novick, 1968, p. 492).

Rasch's philosophy of measurement is compatible with the latent trait theoretical viewpoint which emphasizes that any test situation can be conceived of as an interaction between an individual's ability and the difficulty of the task he is confronted with. The result of this interaction can be expressed as the probability that the individual will successfully meet the challenge. If this type of relationship holds true for any one item in a test and seems theoretically acceptable, one can then consider all items making up that test as providing replications of the same phenomenon. In terms of odds, this relationship can be expressed as:

$$O_{ni} = \Lambda_n E_i \qquad (2.1)$$

where O_{ni} = odds of success for person n on item i

Λ_n = the ability of person n

E_i = the easiness of item i

and where i is any dichotomously scored item.

This formulation reflects the notion of the postulated multiplicative rule. It seems reasonable indeed to assume that if a subject had no ability at all ($A_n = 0$), he should not pass item i no matter how easy that item may be ($O_{ni} = 0$). Similarly, if an item had no easiness, that is, were infinitely difficult ($E_i = 0$), the odds for passing that item should be zero for everyone attempting it, no matter how intelligent that person may be ($O_{ni} = 0$).

Odds can be transformed into probability statements as follows:

$$P_{ni} = O_{ni} / (1 + O_{ni}) = A_n E_i / (1 + A_n E_i) \qquad (2.2)$$

and $$Q_{ni} = 1 / (1 + O_{ni}) = 1 / (1 + A_n E_i) \qquad (2.3)$$

where P_{ni} = the probability of passing item i for person n

and Q_{ni} = the probability of failing item i for person n

and where $P_{ni} + Q_{ni} = 1$

Formulae (2.2) and (2.3) represent the basic statements of the simple logistic test model.

From this general conception of any test situation and with a specific intent to attain objectivity in ability measurement, Rasch explored the similarity between the form of (2.2) and the form of the logistic function because of the interesting properties of that mathematical function.

The logistic function can be represented as:

$$y = f(x) = \frac{e^x}{1 + e^x} \qquad (2.4)$$

where e is a mathematical constant and x varies from minus infinity to plus infinity.

The relationship between (2.2) and (2.4) is established by the following equation:

$$A_n E_i = e^x$$

From this equation, if $x = 0$, $A_n E_i = 1$ and

$$P_{ni} = A_n E_i / (1 + A_n E_i) = \tfrac{1}{2} = 0.5$$

If x = plus infinity, $P_{ni} = 1$, since $e^{\infty} / (1 + e^{\infty}) \simeq 1$

and if x = minus infinity, $P_{ni} = 0$, since $e^{-\infty} / (1 + e^{-\infty}) \simeq 0$

The final link between Rasch's formulation of the model (2.2) and the logistic function (2.4) can be made more explicit by the following expression:

$$x = b_n - d_i$$

where b_n = the ability of person n

and d_i = the difficulty of item i

Then the three preceding statements simply say that when the ability of a person matches the difficulty of an item ($b_n - d_i = 0$ or $x = 0$) the probability for that person to succeed on that item is one half, whereas it is one when the ability is infinite ($b_n - d_i$ = plus infinity) and zero when the difficulty is infinite ($b_n - d_i$ = minus infinity). It can thus be seen that making use of the range of x in the logistic function provides a natural scale for measurement since the ability parameter and the difficulty parameter are allowed to vary from zero to infinity while the probability of passing an item remains between zero and one.

Rasch's model can therefore be also expressed as:

$$P_{ni} = e^x / (1 + e^x) = e^{(b_n - d_i)} \Big/ (1 + e^{(b_n - d_i)}) \qquad (2.5)$$

The link between the right-hand part of (2.5), which is the restricted form of the general logistic test model, and Rasch's formulation of the simple logistic model can be understood by looking at the general logistic test model.

The logistic test model is determined by assuming that item characteristic curves have the form of a logistic cumulative distribution function:

$$P_i(\theta) = \Psi(x) = \Psi(DL_i(\theta))$$

where Ψ represents the logistic function, D is a scaling factor and $L_i(\theta)$ is the parameter space. In this general form,

$$L_i(\theta) = c_i(b_n - d_i)$$

where c_i = discriminating power of item i

d_i = difficulty level of item i

and b_n = ability level n of θ, the underlying trait.

Therefore, the general form of the model can be expressed as:

$$\Psi(x) = \left[1 + e^{-Dc_i(b_n - d_i)}\right]^{-1} \qquad (2.6)$$

Most of the developmental work on the application of the logistic function to mental testing has been done by Birnbaum (Lord and Novick, 1968). When $D = 1$ and $c_i = 1$, (2.6) is equal to (2.5) since:

$$\left[1 + e^{-(b_n - d_i)}\right]^{-1} = \frac{1}{1 + \frac{1}{e^{(b_n - d_i)}}}$$

$$= \frac{1}{\frac{1 + e^{(b_n - d_i)}}{e^{(b_n - d_i)}}}$$

$$= \frac{e^{(b_n - d_i)}}{1 + e^{(b_n - d_i)}}$$

which is Rasch's model, a restricted form of the general model where all

item discriminating powers are assumed to be equal.

Since these two models are latent trait models, it is of interest to briefly examine the common characteristics of this class of models.

Latent trait models

It is not the purpose of this study to provide a complete presentation of latent trait theory. However, since most of the assumptions of Rasch's model can be explained by the main concepts of this theory, a condensed version drawn from Lord and Novick (1968) will be presented here.

The classical linear testing model has been categorized among weak true-score models because of the weakness or generality of its assumptions, and hence, its capability to satisfy most data. This model expresses the relationship between an observed score and a true score in terms of a random error. Item-test regression curves are then derived relating the percent of people passing an item to a base line representing the total test score. However, this type of relationship is meaningless since the total test score is an unreliable measure and the unit of measurement provided by the score scale is specific to the particular test administered. Thus a single item will have differently shaped regressions on different tests measuring the same trait. Most of the practical work so far in ability measurement has relied on weak true-score models. The problems associated with these models have prompted psychometricians to develop stronger models. These new models rely on a general theory of latent traits. Anderson (1959) has

provided a very general formulation of latent trait theory from which specific models derive. This formulation has been restated:

> "Any theory of latent traits supposes that an individual's behavior can be accounted for, to a substantial degree, by defining certain human characteristics called traits, quantitatively estimating the individual's standing on each of these traits, and then using the numerical values obtained to predict or explain performance in relevant situations." (Lord and Novick, 1968, p. 358).

The primary function of these models is thus to relate, in a mathematically rigorous and psychologically meaningful way, a set of observable or manifest variables X_i to another set of unobservable or latent variables θ_i, so that the latter can be inferred from the former.

In the classical model, a true score is inferred from an observed score but there is no way one can relate that true score to the underlying ability θ other than by assuming that they are equivalent. A trait is here understood to represent any psychological dimension of interest including mental abilities (aptitudes and achievements), attitudes, interests, etc. To achieve parsimony, it is hoped that the number of components of θ_i is smaller than that of X_i.

The specificity of each model comes from the nature of the postulated relationship between these two sets of variables. For instance, a model that has been applied to the analysis of sociological data is Lazarsfeld's "latent class model". This model assumes that θ takes a finite set of m values, called classes, and that each individual is classifiable into one of m classes with identical within-class residual distributions. But more relevant to psychological testing are models that assume the trait values to be continuous. In this study, we are interested in very specific types of latent trait models which apply to binary (dichotomously scored) items and assume

that θ is both continuous and unidimensional. Two special cases of the general model are the logistic and the normal ogive models.

Latent trait models share two characteristics. First, they postulate a relationship between the observed variables X_i and the underlying variables θ, $P_i(\theta)$, where $P_i(\theta)$ is the conditional probability of passing item i for a subject at a given ability level of θ.

Because of the stronger assumptions needed and the possibility of testing such a functional relationship, these models are strong true-score models. Instead of deriving item-test score regressions, it is then possible to plot regressions of item scores on θ, the basic ability. These regressions, also called item characteristic functions when θ is multidimensional and item characteristic curves when θ is unidimensional, necessarily remain invariant from one group of examinees to the next provided that θ spans the complete latent space. Hence, any parameter describing the item characteristic function is an invariant item parameter.

> "The item characteristic function is a key concept... for making inferences about unobservable latent traits from the observed item responses. Making such inferences is, as we have said, a basic purpose of mental testing." (Lord and Novick, 1968, p. 360).

Since θ cannot be observed, item characteristic functions cannot be directly observed either. Thus certain assumptions need be made about the shape of these functions. Depending on how one chooses $P_i(\theta)$, different latent trait models follow having different basic properties. For his model, Rasch has postulated that $P_i(\theta)$ had the shape of a cumulative logistic function which is similar to that of the normal ogive model. In both of these models, θ is assumed to be unidimensional.

Lord and Novick (1968) report that the problem of statistically investigating the hypothesis of unidimensionality without specifying the shape of the item characteristic curve (ICC) has not been completely solved. Practically, one is interested in the reasonableness of such an assumption. It should be pointed out that the dimensionality of the latent space is a broader concept than the factor structure of a test. It is therefore possible to have a test that would yield more that one common factor in a factor analysis of inter-item correlations and still be closely approximated by a unidimensional model. This apparent paradox comes from the fact that factor analyses are performed on correlation matrices, and that correlation coefficients depend for their magnitude on the distributional properties of the samples used.

> "The number of common factors in a correlation (or covariance) matrix depends on the type of correlation coefficient (or covariance) used. It also depends on how the item scores are transformed before the correlations are computed. The dimensionality of the complete latent space does not depend on distributional assumptions, nor on a choice of a measure of inter-item correlation, nor on any choice of transformation of the latent variables. Thus the dimensionality of the complete latent space is a more basic concept than is the number of common factors." (Lord and Novick, 1968, p. 382).

The second characteristic shared by all latent trait models is the assumption of local or conditional independence.

> "Local independence means that within any group of examinees all characterized by the same values θ_1, θ_2,..., θ_k, the (conditional) distributions of the item scores are all independent of each other." (Lord and Novick, 1968, p. 361).

Local independence also exists when, for any fixed value of θ, knowing how a given subject performed on item i will not help predict his score on item j.

The homogeneity of a test with respect to its content and the independence of item responses are two desirable characteristics of a

test in the classical as well as in the latent trait approach.

The notion of a model

There are three essential conditions for achieving meaningful measurement: an adequate model, invariant scaling, and objectivity.

Rasch insists that a model is never true or false and deplores the misconceptions too often entertained around the notion of what a model is and what it is for.

> "My point of departure I take in the statement that models are never true and they are not meant to be so. This point may be illustrated by the case of the pendulum. The simplest model in this case is the mathematical pendulum: a heavy point fixed to a weightless string and swinging frictionless in vacuum" (Rasch, 1964).

Since the final goal in ability measurement is to obtain as precise an estimate of an individual's ability as possible by a method that would provide meaningful measurement at a reasonable cost in terms of money, time, and labor, Rasch developed a model to meet those requirements.

For a model to be adequate, it must represent properly the phenomenon of interest. Rasch's formulation of the simple logistic model (2.2) is meant to be a proper representation of the type of response which can be expected when a subject of a given ability level encounters a specific stimulus. The probabilistic nature of that response is thus the first characteristic sought by Rasch. The multiplicative rule ($A_n E_i$) is a second characteristic of Rasch's model and its logical basis was briefly explored in the first section of this chapter. A third characteristic which needs to be looked at a little more is that of the functional relationship between the response and the underlying ability, $P_i(\theta)$.

The central importance of item characteristic curves (ICCs) for latent trait models has been established in the previous section. ICCs represent the key to the notion of invariant scaling.

For any item i from a test administered to a group of subjects of different ability levels $\theta_1, \theta_2, \ldots, \theta_k$, to each ability level corresponds a proportion p_i of subjects who get item i correct which varies from 0 to 1. To infer θ_i from p_i, a function $P_i(\theta)$ must be defined as a correspondence between respective ability scores θ_i and values p_i. However the θ_i are not known since they are the unobservable variables. But even though the θ_i are unknown, a metric can be chosen to represent them. This metric can be selected so that $P_i(\theta)$ is strictly increasing in θ. It thus appears that the choice of the metric for θ will determine the functional form of $P_i(\theta)$. This statement can be logically reversed to mean that the choice of a functional relationship for $P_i(\theta)$ will determine the metric of θ.

Once any specific strictly increasing form has been adopted for item i, one is no longer free to adopt any number as the value of $P_j(\theta)$, nor is one free to adopt any assumption restricting even partially the possible functional forms of any other ICCs. This illustrates the fact that, in general, it is empirically meaningful to assume that any specific model of partially restricted form is either valid or false with respect to a given population of items and hence is subject to empirical confirmation.

> "On the other hand, the assumption that any chosen single item has an item characteristic curve of a specified functional form $P_i(\theta)$ that depends on ability θ is, when considered in isolation, acceptable in principle as a definition of the ability scale of θ values and is not an empirical specification." (Birnbaum, 1968, p. 399).

An ICC is thus a regression of item scores on θ and it has been seen in the previous section that any parameter describing an ICC is an invariant parameter. Each ICC is usually characterized by two parameters, one describing its location and another describing its steepness. The first one is an index of the difficulty level of an item and is defined as the level of ability at which an item discriminates most effectively. The second one is an index of the discriminating power of an item and indicates the quality of an item in the basic sense of the amount of information the item provides about θ. Since these two parameters are contingent upon θ, they should remain invariant across groups of examinees. When contrasted with the classical model parameters, their advantages emerge clearly. The classical model defines item difficulty as the proportion of people that answer an item correctly, and item discrimination as the biserial correlation between item score and test score, two directly sample-bound measures which necessarily vary from group to group.

To summarize, an ICC represents $P_i(\theta)$ which is assumed to be monotonically increasing in θ and specifically defined by its postulated form and a set of parameters $L_i(\theta)$ where $L_i(\theta)$ is a linear function of θ and is equal to $c_i(b_n - d_i)$, as previously seen.

The two most popular models which were developed in the area of mental testing are the logistic test model (2.6) studied by Birnbaum (1968) which assumes that $P_i(\theta)$ has the form of a logistic cumulative function and the normal ogive model studied by Lord (1952, 1953) which assumes that $P_i(\theta)$ has the form of a normal cumulative function.

The normal model can be represented as follows:

$$P_i(\theta) = \Phi(x) = \Phi(c_i(b_n - d_i))$$

$$= (1/(2\pi)^{\frac{1}{2}}) \exp(-u^2/2)\, du \qquad (2.7)$$

The logistic model has already been shown to be:

$$P_i(\theta) = \Psi(x) = \Psi(c_i(b_n - d_i))$$

$$= \frac{e^x}{1 + e^x}$$

The logistic and the normal models, like any other latent trait model, postulate a relationship between observed scores and latent variables which is probabilistic in nature. Such models are also called stochastic response models or quantal response models. It makes sense in psychological testing to adopt such a probabilistic formulation since the observations can never be completely accounted for by the parameters of any model. In relating observations to parameters, one must therefore take this uncertainty into account. The relationship between the two models is very close, the two functions nearly coinciding. As a matter of fact, it has been shown (Lord and Novick, 1968, p. 399) that:

$$\left|\Phi(x) - \Psi[(1.7)\, x]\right| < 0.01 \text{ for all } x$$

Because of the similarity of these two functions, the logistic model and the normal model should represent data equally well. However, the logistic model is mathematically much simpler to work with.

Both models have two item parameters and one ability parameter. But the fact that items are represented by two parameters causes serious problems for the estimation of these parameters.

> "In principle it is impossible to solve for maximum likelihood estimates of all the three sets of parameters because the total number of parameters becomes larger than the possible degrees of freedom. Therefore, an additional assumption has to be made to make the method practically feasible." (Panchapakesan, 1969).

There are two ways to solve this problem. One way is to make distributional assumptions about the ability of the calibrating sample and to estimate only the item parameters. This is acceptable if one is solely interested in the calibration of an instrument but even then, it has the disadvantage of presuming a distribution for θ. The other alternative is to make an assumption about the discriminating powers of items. Because of Rasch's interest in developing individual-centered statistics, he could not adopt the first approach.

> "In his writings, Rasch has emphasized that he is not concerned with distributions over people, but rather with estimation of ability for each person. Thus standard measures such as reliability and validity coefficients, which have meaning only in terms of distribution of scores over people, are of no interest to him." (Lord and Novick, 1968, p. 481).

Furthermore, Rasch has demonstrated a logical link between the desired criterion of measurement, that of objectivity, and the mathematical property of separability, that is, independent estimability of parameters. He has shown that the principle of separability, and consequently of objective measurement, can only be achieved when the model is limited to two parameters.

Therefore, Rasch derived the simple logistic model (2.5) by assuming $c_i = 1$, that is, all items are of equal discriminating power. The robustness of the model with respect to violations of that assumption has been investigated on simulated data. It is one of the objectives of this study to carry out such an investigation on real data. Suffice it to mention at this point that this assumption is not as

unreasonable as it may sound. If item discriminations fall within a certain range, the model treats them as if they were equal. This issue will be dealt with later.

In summary, the model developed by Rasch is assumed to properly represent any measurement situation, makes use of the invariant scaling property of latent trait models, and is the only model capable of yielding truly objective measures. Moreover, as it should be for any model, all of its assumptions are subject to empirical validation.

Assumptions of the model

The simple logistic model relies on three major assumptions, the first two being shared by all latent trait models.

The assumption of local independence states that, given a person's ability, the conditional probability of his getting item j correct is independent of whether he has got item i correct or not, and this is the case for all subjects having the same ability. This is not the same as saying that inter-item correlations computed over the whole sample of subjects are close to zero. But it is the same as saying that if all subjects in that sample have exactly the same ability, these inter-item correlations should be close to zero. In other words, if ability is partialled out, the behavior of the subjects exhibited through their responses should be random provided the uni-trait assumption holds. The fact that variation in abilities tends to increase inter-item correlations or that restriction in the range of abilities tends to reduce these correlations is a consequence of this assumption (Rasch, 1966). In the simple logistic model, inter-item

correlations do not represent intrinsic properties of the items, but are mainly determined by variations in the person parameters.

The uni-trait assumption has been examined in previous sections. A uni-factor test ensures the unidimensionality of the latent space but, as already seen, a one factor structure of the test is not a necessary condition for the latent space to be unidimensional. This assumption adds a few constraints. The model is limited to power tests since a speed factor would violate the uni-trait assumption. The model is also limited to free response items since a guessing factor could have a similar effect.

Finally, the model is restricted to a single item parameter with the assumption that all items have equal discriminating powers of magnitude set as unity.

In this study, only dichotomously scored items will be considered.

Estimation of the parameters

So far the discussion has been limited to the representation of a single item by $P_i(\theta)$. Since both the normal ogive and the logistic test models assume that all items have the same ICC, the form of which is specified by the respective model it can be seen that this would be possible only if items were all of the same difficulty level and of the same discriminating power. No such test exists in practice but this illustrates the notion of a test characteristic function or curve (TCC). The TCC would then have the same form as the ICCs.

It turns out that the overall fit of a model is based on such an expectation, given that the observed responses are conditioned on θ and that the assumption of local independence holds. This latter

assumption which is common to all latent trait models is necessary for estimating an individual's ability as a function of his test score.

In the case of binary items, the response on each item can only be zero or one. The assumption of local independence states that knowing an examinee's score on one item will not help predict his score on any other item. A score on item i, denoted by a_i, is thus related to an ability θ by one of two functions that give the probability of each possible score on item i, that is, $a_i = 1$ or 0. One of these functions is $P_i(\theta)$, the probability of passing item i, that is, of getting a score of 1. The other function is $Q_i(\theta)$, the probability of failing item i, that is, of getting a score of 0:

$$P_i(\theta) = P\ (a_i = 1|\theta) \tag{2.8}$$

$$Q_i(\theta) = 1 - P_i(\theta) = P\ (a_i = 0|\theta) \tag{2.9}$$

Combining (2.8) and (2.9) in the probability distribution of A_i, a single statement is obtained:

$$P\ (A_i = a_i|\theta) = P_i(\theta)^{a_i} \cdot Q_i(\theta)^{(1 - a_i)} \tag{2.10}$$

$$= \begin{cases} P_i(\theta) \text{ if } a_i = 1 \\ Q_i(\theta) \text{ if } a_i = 0 \end{cases}$$

Since $P_i(\theta)$ is chosen to be strictly increasing in θ, the item response a_i will be an indicant and a measure of θ.

To get an unequivocal relation between an ability and a complete response pattern $A = (a_1, \ldots, a_k)$, assuming statistical independence between responses, the probability product form must be computed:

$$\begin{aligned} P(A = a|\theta) &= P\ (A_1 = a_1, \ldots, A_k = a_k|\theta) \\ &= P\ (A_1 = a_1|\theta) \ldots P(A_k = a_k|\theta) \\ &= \prod_i^k P_i(\theta)^{a_i} \cdot Q_i(\theta)^{(1 - a_i)} \end{aligned} \tag{2.11}$$

which is simply the product of the functions relating all item responses to an ability θ.

Since $P_i(\theta)$ is specified differently under different models, as was seen earlier, it should be obvious that (2.11) is much easier to carry out under the logistic model (2.4) than it is under the normal ogive model (2.7).

Rasch (1966) developed the concept of specific objectivity as being a function of the fact that not only can ability estimates be obtained independently from item parameters (and vice-versa) but that a test of the model can be derived which is completely independent of all of the model parameters.

Without giving the elaborate mathematical derivation of that property, the notion of separability can be briefly described by the simple example of a theoretical two-item test.

Consider the situation where individual n takes a test composed of only two items, item i and item j. The assumption of local independence requires that his response to item i, a_{ni}, be independent of a_{nj}, his response to item j. His total test score will be the sum of his item scores, $a_. = a_i + a_j$.

The variable a. can take on the values 0, 1, or 2, with probabilities given by:

$$P(a_. = 0) = P(a_i = 0) \cdot P(a_j = 0)$$

$$P(a_. = 1) = P(a_i = 0) \cdot P(a_j = 1) + P(a_i = 1) \cdot P(a_j = 0)$$

$$P(a_. = 2) = P(a_i = 1) \cdot P(a_j = 1)$$

Substituting for Rasch's specifications of the simple logistic model (2.2 and 2.3):

$$P(a_. = 0) = \frac{1}{1 + A_n E_i} \cdot \frac{1}{1 + A_n E_j}$$

$$P(a_. = 1) = \frac{A_n E_i}{(1 + A_n E_i)(1 + A_n E_j)} + \frac{A_n E_j}{(1 + A_n E_i)(1 + A_n E_j)}$$

$$= \frac{A_n (E_i + E_j)}{(1 + A_n E_i)(1 + A_n E_j)}$$

$$P(a_. = 2) = \frac{A_n^2 E_i E_j}{(1 + A_n E_i)(1 + A_n E_j)}$$

From this, it can be seen that the probability of passing item i given a total score of 1 and an ability A_n is:

$$P(a_i = 1 | a_. = 1, A_n) = \frac{P(a_i = 1)}{P(a_. = 1)}$$

$$= \frac{\dfrac{A_n E_i}{(1 + A_n E_i)(1 + A_n E_j)}}{\dfrac{A_n (E_i + E_j)}{(1 + A_n E_i)(1 + A_n E_j)}}$$

$$= \frac{E_i}{E_i + E_j}$$

which is a demonstration of the fact that A and E can be estimated separately since one results in having a probability statement relating the parameters E_i and E_j to the data so that now the E_i and E_j can be estimated without even considering the A_n parameter.

It has been seen previously that the general logistic model retains two item parameters:

$$\Psi(x) = P_i(0) = \frac{e^{c_i (b_n - d_i)}}{1 + e^{c_i (b_n - d_i)}} \qquad (2.12)$$

Since c_i interacts with both b_n and d_i, it becomes much simpler for estimation purposes to get rid of that parameter, much the same way one would postulate the absence of interaction in an ANOVA factorial design. One way of doing so is thus to assume equality of item discriminations and to set $c_i = 1$. The soundness and implications of that assumption will be examined later on.

The restricted form of the model is thus obtained (2.5):

$$P_i(\theta) = \frac{e^{(b_n - d_i)}}{1 + e^{(b_n - d_i)}}$$

But the model is also represented as follows:

$$P_i(\theta) = \frac{e^{(b_n + d_i)}}{1 + e^{(b_n + d_i)}} \qquad (2.13)$$

The only difference between (2.5) and (2.13) is in the sign of the parameter d_i. In the first case d_i represents the difficulty of item i whereas it reflects the easiness of item i in the second form. In fact, in the relation $A_n E_i = e^{(b_n - d_i)}$,

$$A_n = e^{b_n} \text{ and } E_i = 1/e^{d_i}$$

whereas in $A_n E_i = e^{(b_n + d_i)}$,

$$A_n = e^{b_n} \text{ and } E_i = e^{d_i}$$

A log transformation is used to ease the process of estimating b_n and d_i.

Since $A_n E_i = e^{(b_n + d_i)}$,

$\log (A_n E_i) = \log \left[e^{(b_n + d_i)}\right]$, and

$\log A_n + \log E_i = b_n + d_i$

thus, $b_n = \log A_n$ (log ability)

and $d_i = \log E_i$ (log easiness)

In its log form, the model becomes:

$$P_i(\theta) = \frac{\exp (b_n + d_i)}{1 + \exp (b_n + d_i)} \tag{2.14}$$

$$Q_i(\theta) = \frac{1}{1 + \exp (b_n + d_i)} \tag{2.15}$$

As mentioned earlier (2.10), it is useful to express the model in a more compact form which integrates $P_i(\theta)$ and $Q_i(\theta)$ by the use of the binomial function.

For a person n, the probability of a given response to one item i is thus:

$$P (A_{ni} = a_{ni} | \theta) = P_i(\theta)^{a_{ni}} \cdot Q_i(\theta)^{(1 - a_{ni})}$$

$$= \frac{(A_n E_i)^{a_{ni}}}{1 + A_n E_i}$$

$$= \frac{\exp (a_{ni} (b_n + d_i))}{1 + \exp (b_n + d_i)} \tag{2.16}$$

But (2.16) expresses the probability of a response to only one item by just one person. Assuming conditional independence between all item responses, the probability of an individual test score (k items) is simply the product of single item score probabilities:

$$P(A_n) = \frac{\prod_{i=1}^{k} [\exp(a_{ni}(b_n + d_i))]}{\prod_{i=1}^{k} [1 + \exp(b_n + d_i)]} \qquad (2.17)$$

However (2.17) represents the probability of only one score pattern. We know that a total score R_n may be made up of many different score patterns. The probabilities of all possible response vectors A_n with exactly R correct responses must therefore be summed up to get the probability of R_n. Thus the probability of a given test score for person n with ability b_n is:

$$P(R_n) = \sum_{r=1}^{R} P(A_n) \qquad (2.18)$$

which is simply the sum of the values obtained for each pattern by formula (2.17).

Finally, an expression for the total probability of obtaining the observed responses A of N persons to K items must be derived. Assuming independence among the responses of different persons,

$$P(A) = \prod_{n=1}^{N} P(A_n) \qquad (2.19)$$

which is simply the product of the values obtained for each pattern by formula (2.17).

> "An important consequence of this model is that the number of correct responses to a given set of items is a sufficient statistic for estimating person ability. This score is the only information needed from the data to make the ability estimate. Therefore, we need only estimate an ability for each possible score. Any person who gets a certain score will be estimated to have the ability associated with that score. All persons who get the same score will be estimated to have the same ability." (Wright and Panchapakesan, 1969).

Since the score pattern of an individual is of no interest, there will be as many estimates of the ability parameter b_n as there are score groups j. Therefore (2.16) can be rewritten in terms of score groups:

$$P(A_{ni} = a_{ni}|\theta) = \frac{\exp\ (a_{ni}(b_j + d_i))}{1 + \exp\ (b_j + d_i)} \qquad (2.20)$$

where b_j is the ability of all persons who get a score of j on the test.

Wright and Panchapakesan (1969) described two methods of estimation, one using unweighted least squares, the other using maximum likelihood. The most popular and certainly the best estimating procedure is the maximum likelihood method (Cramer, 1962) since it reaches more precise estimates of the model parameters and better approximations to the standard errors of estimate.

> "However, when the calibration sample is large, and the ability range of the sample is wider than the easiness range of the item parameters, then the item estimates obtained by LOG (unweighted least squares) are equivalent to the estimates obtained by MAX (maximum likelihood)." (Wright and Panchapakesan, 1969).

Both procedures, as described by Wright and Panchapakesan (1969), will be summarized here since this study makes use of the computer program developed by these authors.

The log method of estimation uses the observed proportion of correct responses within a particular score group j (a_{ji}/r_j) as an estimate of p_{ji}, the probability of obtaining a right answer for any person in score group j to an item of easiness $E_i = \exp\ d_i$. Therefore,

$$p_{ji} = \frac{\exp\ (b_j + d_i)}{1 + \exp\ (b_j + d_i)} \qquad (2.21)$$

where b_j = the ability of subjects in score group j

r_j = the number of subjects in score group j

a_{ji} = the number of subjects in score group j who get item i correct

If $p_{ji} = a_{ji}/r_j$, then $q_{ji} = r_j - a_{ji}/r_j$

$$\text{and } \frac{p_{ji}}{q_{ji}} = \frac{a_{ji}}{r_j - a_{ji}} \simeq \exp(b_j + d_i) \tag{2.22}$$

$$\text{since } q_{ji} = \frac{1}{1 + \exp(b_j + d_i)} \tag{2.23}$$

Therefore, by setting $t_{ji} = \log(p_{ji}/q_{ji})$,

$$t_{ji} = \log(a_{ji}/(r_j - a_{ji})) \simeq b_j + d_i \tag{2.24}$$

$$\text{so } t_{ji} = b_j^* + d_i^* \tag{2.25}$$

where b_j^* = estimate of b_j

and d_i^* = estimate of d_i

which gives the estimation equations

$$d_i^* - d_\cdot^* = t_{\cdot i} - t_{\cdot\cdot} \tag{2.26}$$

where $d_\cdot^* = (1/k) \sum_{i=1}^{k} d_i^*$, that is, the average of all item easiness estimates.

The indeterminacy in the product $A_n E_i$ (multiplying ability by any factor does not change that product provided that easiness is divided by the same factor), which is reflected in the scale of easiness can be removed by setting $d_\cdot^* = 0$.

$$\text{Then, } \log E_i^* = d_i^* = t_{\cdot i} - t_{\cdot\cdot} \tag{2.27}$$

where $t._i = (1/k-1) \sum_{j=1}^{k-1} t_{ji}$

and $t.. = (1/k) \sum_{i=1}^{k} t._i$

The estimation equation for ability thus becomes:

$$\log A_j^* = b_j^* = t_j. - t.. \qquad (2.28)$$

Equations (2.27) and (2.28) are the fundamental equations for the log procedure of estimation.

Since a_{ji} has a binomial distribution, its variance will be given by:

$$V(a_{ji}) = r_j p_{ji} (1 - p_{ji})$$

The variance of t_{ji} can thus be approximated from:

$$V(t_{ji}) \simeq (\partial t_{ji}/\partial a_{ji})^2 V(a_{ji})$$

$$\simeq 1/r_j p_{ji} (1 - p_{ji})$$

or $V^* (t_{ji}) = 1/r_j p_{ji}^* (1 - p_{ji}^*)$

where $p_{ji}^* = \dfrac{\exp (b_j^* + d_i^*)}{1 + \exp (b_j^* + d_i^*)}$

and $(\partial t_{ji}/\partial a_{ji})$ is the partial derivative of t_{ji} with respect to a_{ji} and equals:

$$1/r_j p_{ji}^* (1 - p_{ji}^*)$$

The variance of d_i^* is obtained from (2.27):

$$V(d_i^*) = V (t._i - t..)$$

Assuming that the t_{ji}'s are independent of each other,

$$V(d_i^*) \simeq V (t._i)$$

so

$$V^{*}(d_i^{*}) = (1/(k-1))^{2} \sum_{j=1}^{k-1} V(t_{ji}) \tag{2.29}$$

The variance of the ability estimate is:

$$V^{*}(b_j^{*}) = (1/k^{2}) \sum_{i=1}^{k} V(t_{ji}) \tag{2.30}$$

The square root of (2.29) and (2.30) provides the standard errors of the easiness estimates and of the ability estimates respectively.

For the maximum likelihood procedure, one assumes that the observed data is the most likely occurrence so that parameters are estimated to maximize the likelihood of obtaining the sample of observations. (Wright and Panchapakesan, 1969, p. 34).

The equations are:

$$a_i = \sum_{j=1}^{k-1} (r_j \exp(b_j^{*} + d_i^{*}) / (1 + \exp(b_j^{*} + d_i^{*}))) \tag{2.31}$$

$$i = 1,2, \ldots, k$$

$$j = \sum_{i=1}^{k} (\exp(b_j^{*} + d_i^{*}) / (1 + \exp(b_j^{*} + d_i^{*}))) \tag{2.32}$$

$$j = 1,2, \ldots, k-1$$

where a_i = the number of persons who get item i correct

j = the score group

r_j = the number of persons in score group j

and the log likelihood is

$$\sum_{i=1}^{k} \sum_{j=1}^{k-1} a_{ji}(b_j + d_i) - \sum_{i=1}^{k} \sum_{j=1}^{k-1} \log(1 + \exp(b_j + d_i))$$

The estimates b_j^* and d_i^* are computed from the implicit equations (2.31) and (2.32) which are treated as two independent sets and solved accordingly. But since they are implicit in d_i^* and b_j^*, they cannot be solved directly. Wright and Panchapakesan (1969) use the Newton-Raphson procedure to obtain estimates for the unknown parameters. This iterative procedure is described by these authors and will not be repeated here.

An approximation of a standard error for item estimates is given by:

$$V^* (d_i^*) \simeq 1/ \sum_{j=1}^{k-1} \left[r_j \exp (b_j^* + d_i^*) / (1 + \exp (b_j^* + d_i^*))^2\right]$$

and an approximation of a standard error for ability estimates by:

$$V^* (b_j^*) \simeq 1/ (C (b_j^*) \exp (b_j^*)) + (1/C^2 (b_j^*)) \cdot \sum_{i=1}^{k} \left[V (d_i) \left[\exp (d_i) / (1 + \exp (d_i + b_j^*))^2\right]^2\right]$$

where $C (b_j^*) = \sum_{i=1}^{k} \left[\exp (d_i) / (1 + \exp (b_j^* + d_i))^2\right]$

Fit of the model

Each a_{ji}, the number of right answers to item i obtained in score group j, has a binomial distribution with parameters p_{ji}, the probability of making a correct response and r_{ji}, the number of persons with a score j.

Wright and Panchapakesan (1969) formed a standard deviate and used it as a test of item fit:

$$y_{ji} = (a_{ji} - E (a_{ji})) / (V (a_{ji}))^{\frac{1}{2}} \qquad (2.33)$$

where $E(a_{ji}) \simeq r_j p^*_{ji} = r_j \exp(b^*_j + d^*_i) / (1 + \exp(b^*_j + d^*_i))$

and $V(a_{ji}) \simeq r_j p^*_{ji} (1 - p^*_{ji})$

An approximate χ^2 statistic can thus be obtained for each item by summing y^2_{ji} over the score groups to give:

$$\chi^2_i = \sum_{\substack{j=1 \\ r_j \neq 0}}^{k-1} y^2_{ji}$$

with degrees of freedom = m - 1

where m = the number of non zero score groups.

It is suggested not to mechanically delete all items for which χ^2_i is significant at some level, since the statistic is only approximate, but rather to examine in detail items for which χ^2_i is large. This item statistic is based on the fact that y_{ji} will have an approximately unit normal distribution if item i fits the model and the score group r is large enough.

An overall test statistic can also be obtained by summing the squared unit normal deviates over the entire matrix Y:

$$\chi^2 = \sum_{i=1}^{k} \sum_{\substack{j=1 \\ r_j \neq 0}}^{k-1} y^2_{ji}$$

with degrees of freedom = (k-1) (m-1)

The degrees of freedom come from the number of observations in the data matrix (k x m) minus the m constraints on the score margins (since $\sum_{i=1}^{k} a_{ji} = jr_j$) minus the degrees of freedom used to estimate (k-1) item parameters.

This is the test of fit that we used for our analyses along with an overall likelihood ratio test developed by the same authors.

Because the actual values of the item chi-square and the overall chi-square test statistics are not reported by the computer program used for this study, we shall examine instead the values of the item mean square (item chi-square statistic divided by its degrees of freedom) and the overall mean square (total test chi-square statistic divided by its degrees of freedom).

However, other approaches have been described. For instance, Keesling (1969) suggested the use of a data analysis method instead of a statistical approach. He described his graphical item analysis method but did not provide any validation data. In the more traditional stream, Andersen (1973) presented a goodness of fit test for each item based on a comparison between the within-score groups estimates and the overall estimates of item difficulties using a conditional maximum likelihood ratio.

In this study, we shall examine the problem of fit by making use of both approaches, the descriptive and the statistical.

CHAPTER III

PREVIOUS STUDIES OF FIT

Georg Rasch's book published in 1960 represents a starting point for systematic research on the simple logistic model applied to measurement. Even though there were a few articles on the Rasch model appearing in America between 1960 and 1967, it was not until the 1967 ETS Invitational Conference on Testing that the interest of the American measurement community was stirred. The paper presented at that conference by Professor Benjamin Wright has served to popularize the Rasch model more than any other work. Rentz and Bashaw (1975) estimate that research dated since the Wright paper now numbers well over 300 papers.

Rasch's work (1960, 1961, 1964, 1966 (a), 1966 (b)) focuses on the development of a theory of objectivity in measurement. His book and articles reflect a strong desire to convince others that truly objective measurements are now possible. The structure of the model, its characteristics and assumptions, its applicability and limitations are fully explored and the necessary mathematical demonstrations are provided. The soundness of the theory is well recognized by American leading psychometricians as can be seen in Lord and Novick (1968). However, Rasch himself admitted that the test of fit of the model and the estimation of its parameters needed more refinement.

> "This state of affairs leaves a great deal of freedom to the statistician with the risk of the model-testing being at the mercy of his personal preferences". (Rasch, 1964).

A description of different statistical procedures thus appeared

in the literature along with specifications for computer programming (Wright, 1968, Panchapakesan, 1969, Wright and Panchapakesan, 1969, Wright and Mead, 1975, Wright and Douglas, 1975 (b), Bramble, 1969, Keesling, 1969 (a), 1969 (b), Andersen, 1970, 1972, 1973). The scaling properties of the model were also analyzed (Vogt, 1971, Brink, 1972).

However, studies of model-data fit were initiated in America with Wright's paper (1968). Since Chapter II explored the mathematical rationale and statistical properties of the Rasch model, we shall restrict the present chapter to a brief review of some of the papers which constitute milestones in the general area of model-data fit.

The purpose of Wright's first paper (1968) was to demonstrate that sample-free test calibration and person measurement was indeed possible. To do so, Wright used the responses of 976 beginning law students to 48 reading comprehension items on the Law School Admission Test. In order to examine the dependence of test calibration on the abilities of these law students, he put the 325 students who did worst on the test into a dumb group and the 303 who did best into a smart group. There were 10 points difference between the smartest of the dumb group and the dumbest of the smart group. Obviously any traditional person-bound calibration (percentile ability measures) based on one group had to be incomparable with one based on the other group since these two calibrations would not even overlap. Using such an exaggerated situation, Wright showed how the Rasch model could be sample-free. The two calibration curves obtained with the simple logistic model nearly coincided thus supporting the notion that this new way of calibrating tests was free from the effects of the ability distribution of the persons used for the calibration.

This procedure is basically the one that shall be adopted for an important part of our investigations (Chapter VI). The second aspect examined by Wright was that of item-free person measurement. He divided the 48 items on the original test into two subtests of 24 items each with no items in common between them. The two subtests were made as different as possible, the 24 easiest items being used to make an easy test and the 24 hardest items to make a hard test. He then proceeded to illustrate that the ability estimates based on the easy test were statistically equivalent to those based on the hard test by splitting the score each student earned on the whole test into a subscore on the easy test and a subscore on the hard test and converting these scores into ability measures on a common scale by the use of the three calibration curves computed from the whole test and the two subtests. Wright found that the distribution of ability differences (contrary to score differences) was nicely situated around zero which is a demonstration of the fact that person measurement can be independent of item selection. But to take errors of measurement into account, Wright standardized the differences in ability estimates. The difference between the easy test and hard test ability estimates was divided by the measurement error of this difference to produce a standardized difference. It was the distribution of these standardized differences that indicated whether or not the two ability estimates were statistically equivalent. If they were, this standardized variable would have a mean of zero and a standard deviation of one. Wright's results showed a mean of 0.003 and a standard deviation of 1.014.

The first extensive study of model-data fit to be conducted on the Rasch model following Wright's article was Panchapakesan's

dissertation (1969). Since our investigations tend to expand from Panchapakesan's work, her approach and results will be examined more fully than other studies. We shall explore more particularly her investigations related to item discrimination and guessing.

The most often challenged assumption of the Rasch model is that of the homogeneity of item discriminations. The objection that all items are not equally discriminating is certainly valid. The contention of the model is that the item analysis is only applicable to items of similar discrimination. Is such a model realistic? Panchapakesan (1969) examined this problem in a systematic manner. She first attempted to establish a range of discrimination values within which items would be treated as homogeneous by the simple logistic model. She then looked at the effect the variation in discrimination had on the measurement.

In her first series of investigations with simulated data, she used two statistical criteria for detecting bad items, those items which differed in discrimination from the rest of the items. These statistics were the item chi-square test described in Chapter II and the standardized difference of the estimated slope from unity which will be described in Chapter V.

She first simulated a 20 item test with item 10 as the "bad" item. The response data for good items were simulated according to the model and, for the misfitting item, c, the item discrimination was varied between 0.0 and 0.8. Values of c greater than unity were not considered because the effect of an item with discrimination c' greater than unity is the same as an item with a discrimination $c = 1/c'$. For each value of c, the data were simulated for different sample sizes

(from 100 to 2000). The items were assumed to be uniformly distributed in easiness in the range plus or minus 2. The "bad" item was chosen approximately at the middle of the easiness range where it is best estimated since most of the persons in the sample contribute to its estimation.

The two criteria did a good job of picking out item 10 for all values of c from 0.0 to 0.7. However, when $c = 0.8$, item 10 seemed almost indistinguishable from the rest of the items. From this empirical investigation, Panchapakesan concluded that when the discrimination varies between 0.8 and 1.25, we cannot detect any departure of the item from the model. But for item discriminations smaller than 0.4 and greater than 2.5, the criteria suggested could clearly identify the "bad" items. Thus, items of discrimination 0.8 - 1.25 seem to be treated as homogeneous by the model.

For the remaining simulations, Panchapakesan used $c = 0.8$ and $c = 0.4$ as the lower limit for the range of the designs.

Her second simulation was aimed at generalizing the results obtained with a 20 item test. Since it should be even easier to identify the "bad" item in a longer test, she simulated test lengths of 5 and 10 items. A middle item, item 3 for $K = 5$ and item 5 for $K = 10$, was chosen to be the divergent item. The sample size was set at 500.

The evidence obtained with 10 item tests confirmed the results: an item with $c = 0.8$ is indistinguishable from the rest of the items while an item with $c = 0.4$ is readily identifiable. No conclusive statement about the selected discrimination levels could be made from the evidence obtained for 5 item tests.

In a third simulation, Panchapakesan increased the number of "bad" items on the basis of the fact that there normally are several discrepant items in a set of uncalibrated items. She therefore chose 5 of 20 items (25%) to be "bad" items. The five items were selected at random and c was set at 0.8 and 0.4 for sample sizes of 500 and 1000.

Here again she found that items with c = .8 could not be consistently identified while items with c = .4 were certainly the worst items. She found these results heartening because even at sample sizes of 1000, which are not excessive for test calibrations, the proposed analysis could identify an homogeneous set of items for the purpose of calibration and measurement.

In the above examples, c was set at 1.0 for all "good" items. A fourth simulation was then designed to consider the more realistic case when all items vary in discrimination within a certain range. Data were simulated for two ranges of discrimination for test lengths of 20 items. In one set c varied between 0.8 and 1.2 and in the other set c varied between 0.4 and 1.6. The sample sizes selected were 500 and 1000. There was no relationship between the easiness and discrimination of the item, the variation in discrimination being at random.

In this situation, the results were not absolutely clear-cut. For the smaller range, the data did fit the model. In the other case, although the generating discrimination parameters were outside the range which is considered indistinguishable, not all of the items outside that range were identifiable. However, they were the worst items when both the criteria of mean square fit and standardized

difference of the slope were considered. Thus, Panchapakesan concluded that even if the basic nucleus of "good" items does not have exactly the same discrimination, the analysis is sensitive enough to enable us to pick out items which are divergent in discrimination. "The consistency in the results obtained is a sign that in practical problems the proposed analysis can be utilized effectively to select sets of homogeneous items". (Panchapakesan, 1969, p. 88).

Finally, she simulated one "bad" and five "bad" items among a set of 20 items which varied in discrimination in the range 0.8 - 1.2 for sample sizes of 1000. She found again that the "bad" items could be clearly identified in both situations.

In her second series of investigations, Panchapakesan studied the effect of variation in discrimination on measurement. Since items with discriminations in the range 0.8 and 1.25 could not be identified, would this dispersion be significant from the point of view of measurement or could discriminations within this range be considered similar? She attempted to answer this question by simulating subjects of different abilities taking the same test repeatedly. The ability range was set at $\pm$ 2.5 and steps of 0.25 were chosen. Data were simulated for two sets of items. In case I the minimum discrimination was picked so that during calibration the items in the set could not be differentiated on the basis of their discrimination. Case II included the maximum variation in discrimination that is found in actual test data. The items were distributed such that for half of them the discriminations for the other half were set so that the product of all the discriminations was unity. For instance, if $c_{min} = 0.8$, $c_{max} = 1.25$ for the first case, and if $c_{min} = 0.4$, $c_{max} = 2.50$ for the second case.

For the simulation, a 20 item test was chosen with easiness varying uniformly in the range $\pm$ 2. Two arrangements were used for the variation in discrimination. In the first arrangement, the variation in discrimination was at random with respect to the easiness of the item while in the second arrangement the variation in discrimination was such that the discrimination increased with the difficulty of the item.

To evaluate the effect of variation in discrimination on measurement, Panchapakesan looked at the bias and the standardized bias in the measurement. As a measure of the efficiency of the measurement, she plotted the standard deviation and standard error against the ability.

For the random case, she noted that the bias was significant at some values of ability, being greater for the range 0.4 - 2.5 than for the range 0.8 - 1.25. This compatibility between simulation at the calibration stage and simulation at the measurement stage was considered very satisfying, giving added confidence to the procedure used.

For the simulation where the discrimination and the difficulty of the item were correlated, she found a positive bias throughout the entire range. She thus concluded that when the discrimination of items increases with difficulty, a person would be estimated to be more able than he actually is on the basis of the simple logistic model. When ordering was reversed so that the easier the item the greater the discrimination, there was a negative bias throughout the entire range, that is, a subject would be estimated as being less able than he actually is. This should not be too disturbing since in a pool of items such a systematic variation of discrimination with respect to

easiness of the item would not ordinarily be found. In case I the standardized bias was less than .5 throughout the entire range whereas it was quite large in case II, in some cases exceeding one.

For the random case, the standardized bias was less than unity everywhere. This means that even for case II the bias was less than the error of measurement. Panchapakesan concluded that in practical applications the model is robust even when the condition of equal discrimination is not met. Therefore, according to Panchapakesan, when the discrimination of the items varies outside the range specified by the simple logistic model, that model can still be used to make measurements provided that the variation in discrimination is not too extreme. The interval of 1.0 plus or minus 0.20 defined as the amount of slope deviation tolerated by the model is also that established by Rentz (1975).

Let us now turn our attention to Panchapakesan's investigations of the effect of guessing on the Rasch model.

The simple logistic model is only valid for free response items. For large scale testing, multiple choice items are most often used because of ease in scoring. In this case, a subject's response is not only influenced by his knowledge but also by some personality variables. Since guessing is an important factor in multiple choice tests many attempts have been made to control it. However, none of the approaches have been entirely successful. A model commonly used to minimize the effect of guessing is the random guessing model. According to this model if a subject knows the answer to an item he gets it right, and if he does not know the answer he guesses at random among the possible alternatives. In such a model, it is

assumed that all alternatives will look equally attractive to all subjects who are not sure of the correct answer. We know that this is not the case. A more able subject will be able to eliminate some of the distractors, that is, he will guess "intelligently" whereas a less able subject may not be able to eliminate any of the distractors.

In her investigation with simulated data, Panchapakesan used a model for guessing which makes a provision for "intelligent" guessing. That model will be described in Chapter V. She then studied the effect of guessing on item calibration and on measurement.

To explore the effect of guessing on item calibration, Panchapakesan (1969) looked at two cases. In the first case, guessing was assumed to be operating in only one item. The item chosen was of medium difficulty and the responses for the rest of the items were generated according to the simple logistic model. At that stage, her motivation was to see whether the item chi-square was sensitive to an item in which guessing was taking place. She simulated data for a test of 20 items, all items having the same discrimination, varying uniformly in easiness between the range + 2 and - 2. The sample size selected was 5000. The population was uniformly distributed between the range - 3 and -0.5. This range was determined by the threshold value of p specified in her model at which guessing begins to be effective. She used two kinds of tests, a true and false test ($m = 2$) and a multiple choice test ($m = 5$). She found that because of guessing, the chosen item was estimated to be much easier than its generating value both for $m = 2$ and $m = 5$. This should be expected since items appear easier than they are when guessing is permitted. However, the easiness standardized difference (between generating and estimated easiness

values) cannot be used in real data because for real data there is no way of knowing the truth. Panchapakesan looked at the p value for the item chi square and found that it was highly significant for both $m = 2$ and $m = 5$. She thus concluded that in real data a large item chi-square should indicate the presence of guessing in a particular item.

In the second case guessing was allowed for the entire test. The number of items and the range in easiness and discrimination remained as in the first case. She was here interested in seeing how the calibration of the items improved as the ability of the sample increased. The simulations were limited to the case $m = 5$ and the range of the ability of the standardizing sample was varied. The sample size for all the simulations was 1000. Four ranges of ability were considered, - 2 to + 2, - 1 to + 3, 0 to + 4, and + 1 to + 5. Panchapakesan found that the estimates for the ability range + 1 to + 5 did not differ noticeably from the estimates where there was no guessing. From this information, she set up guidelines for minimizing the effect of guessing on item calibration. We shall describe these guidelines in Chapter V when we look at our data. Panchapakesan's main conclusion was that if we can get samples of sufficient ability for calibrating items, we do not have to worry about guessing. In other words, if the average ability of the calibrating sample is greater than the average difficulty of the test, it is reasonable to assume that very few subjects will resort to guessing in responding to test items.

To examine the effect of guessing on measurement, Panchapakesan simulated subjects in the ability range - 2.5 and + 2.5, at intervals of 0.25, allowed to guess while responding to all items. At each ability

level, a subject was replicated 225 times. The easiness of the items varied uniformly in the range + 2 to - 2. She looked at three tests of lengths 10 items, 20 items and 40 items with $m = 2$ and $m = 5$ for each of them. She found that the measurement was free from guessing at values of ability of approximately + 1.25 for $m = 2$ and -0.25 for $m = 5$. She thus concluded that ability measurements which are free from guessing can in fact be made from multiple choice tests if the measurement is restricted to a region satisfying the lower limit given by r*. The formula for computing r* shall be presented in Chapter V.

Another factor which would violate the assumptions of the simple logistic model is speed. The model does not hold for speeded tests since the response of a subject to an item cannot depend on his ability if the subject does not have time to read the item. Therefore, the model cannot apply if the subjects are unable to finish the test in the time allowed. For example, Rasch found that only the results of two of the four subtests he analyzed could be satisfactorily represented by the model. Further analysis showed that the failure of the other two tests to fit the model was due to the effect of time on performance (Rasch, 1964 and 1966(b)). Panchapakesan's work did not consider speeded or partially speeded tests.

The three major application areas for which the potential usefulness of the Rasch model was explored are test design and item selection (Panchapakesan, 1969, Douglas, 1975), self-tailored testing (Wright and Douglas, 1975(a)), and equating of test forms (Cartledge, 1974, Rentz, 1975, Rentz and Bashaw, 1975).

These studies indicate clearly how to proceed in those areas with the use of the simple logistic model. They shall not be reviewed extensively here since the purpose of our work is not to provide this kind of information but to indicate how reliable the various indices of fit described in the literature are and what particular type of fit is needed for any given application area. In Chapter V, we shall refer mostly to Panchapakesan's indices of fit at the item level whereas, in Chapter VI, the work of Rentz and Bashaw (1975) on the invariance properties of ability and easiness parameter estimates will be examined in more detail.

CHAPTER IV

TEST, SAMPLE, AND PROCEDURES

Introduction

The purpose of this study is to explore the concept of model-data fit using the Rasch model applied to some real data. The notion of fit between a mathematical model and actual observations is always a complex matter. When there is a lack of fit between a model and some data, one never knows for sure how much of this misfit is due to the inadequacy of the model in representing the data and how much is attributable to the error component present in any set of data. One way to solve such a dilemma is to apply the model repeatedly to similar kinds of data and, progressively, to span the whole universe to which those data relate. But this is a costly enterprise. Fortunately, the use of simulation has facilitated the process and made such a goal reachable at a reasonable cost. We are interested in knowing how well the simple logistic model can represent real test data. We shall now present briefly the means that shall be used in this study to achieve our purpose.

The test

In order to fully explore the concept of fit, a test must be chosen that has at least the following features. First, the size of the sample must be large so that analyses could be performed on sub-samples of still adequate sizes. Second, the sample should be

heterogeneous with respect to ability and other individual characteristics to allow for investigation of as many hypotheses of misfit as possible. Third, the test itself should measure various types of abilities so that it becomes possible to check for varying degrees of fit of a given measurement model to different kinds of abilities. Fourth, the test should have items of varying degrees of difficulty to match the levels of ability of the sample.

These requirements were found in the Medical College Admission Test (MCAT) and, after reaching an agreement with the American Association of Medical Colleges (AAMC) responsible for the MCAT program, actual data were secured and used for our investigations.

The MCAT is administered twice yearly, in May and October, at centers throughout the United States and at foreign centers throughout the world. The overall format of the test is the same but different forms are used for each administration. Test administration and scoring is handled by the Psychological Corporation. As of January 1970, the Division of Educational Measurement and Research of the AAMC assumed responsibility for all test reports to schools and to examinees and for all research related to the MCAT program.

The overall test is made up of four subtests of different length and testing time:

Subtest	Number of items	Testing time
Verbal Ability	75	20 min.
Quantitative Ability	50	45 min.
General Information	75	25 min.
Science	86	60 min.

All four subtests are power tests. Each is designed so that nearly all applicants will have an opportunity to respond to all of the questions. The time limits are used primarily to achieve administrative uniformity and not to speed responses (Sedlacek, 1967). The MCAT consists entirely of four-option, multiple-choice items. Within each subtest, the questions are ordered from easiest to most difficult. They are not grouped by subject matter content. Raw scores are obtained as simple counts of the number of items answered correctly, with no correction for guessing. Scaled score transformations are based on the performance of some 12,500 individuals who took the MCAT in 1951 (Reference Group).

The Verbal Ability subtest comprises 75 items, 30 based on synonyms, 25 on antonyms, and 20 on verbal analogies. Its purpose is to measure vocabulary strength and ability to perceive verbal relationships, qualities presumed to be indicative of general ability to handle postgraduate study.

The Quantitative Ability subtest comprises 50 items based on arithmetic, elementary algebra, and geometry, but its primary purpose is to assess the ability to reason with numerical and quantitative concepts rather than to test for specific mathematical knowledge or achievement. Along with Verbal Ability, it is presumed to be indicative of general academic aptitude.

The General Information subtest consists of 75 items and is designed to give an indication of the applicant's breath of knowledge in such fields as history, government, political science, economics, geography, sociology, anthropology, psychology, literature, philosophy, art, music, and even sports.

The Science subtest consists of 86 items, 50 percent dealing with chemistry, 35 percent with biology and 15 percent with physics. Understanding of functions is stressed rather than knowledge of taxonomic details in biology, and understanding of principles and problem-solving rather than recall of isolated bits of information in physics and chemistry (Erdmann et al., 1971).

The overall purpose of the MCAT could be summarized as follows:

> "The test is intended to provide admissions committees of medical schools with information about certain abilities of their applicants which may be used in conjunction with other information, such as that gathered from application forms, undergraduate records, recommendations, and interviews, in making decisions about acceptance or rejection." (Erdmann et al., 1971).

Our analyses were performed on the MCAT administered in May of 1972, which is referred to as Series 49 by the AAMC. For our purposes, the General Information subtest was not used. The reasons for excluding this subtest were numerous. First, in its revised version of the MCAT, the A.A.M.C. does not intend to retain such a section. Second, it has not been used that much by medical schools in their selection process. Third, the variety of tests needed for this study was already sufficient. Fourth, it was not felt that the extra cost involved would have provided interesting payoffs in terms of additional valuable data. Thus, the three subtests selected were Verbal Ability Form U, Quantitative Ability Form R, and Science Form F.

The sample

On each administration of the MCAT, the candidates are asked to fill out a questionnaire. The data are then summarized and published by the Division of Educational Measurement and Research of the AAMC once a year. For 1972, the characteristics of the 51,695 subjects

TABLE 1

CHARACTERISTICS OF MCAT
EXAMINEES IN 1972

CHARACTERISTIC	TOTAL	PERCENTAGE NON-REPEATING	PERCENTAGE REPEATING
SEX:			
Male	82	81	84
Female	18	19	16
COLLEGE STATUS:			
Sophomore	5	6	0
Junior	40	51	10
Senior	28	22	46
College graduate	27	21	44
UNDERGRADUATE MAJOR:			
Biological Sciences	48	47	51
Humanities	4	4	4
Physical Sciences	22	23	19
Social Sciences	12	12	13
Premedical	8	8	7
Other	7	7	6
REGION:			
Northeast	28	28	27
Southeast	12	12	14
North Central	24	24	23
South Central	14	13	15
Far West	15	14	15
Canada	6	6	3
Foreign	2	2	2

who took the test are presented in Table 1. For that year, repeating examinees were 26.1% of the total examinee group. There are no data available individually for the May (Series 49) and the October (Series 50) administration of the test but the percentages do not vary a great deal over the years between both sessions and between a single session

and the year totals.

The data gathered from the May administration of the test (Series 49) were provided to us on tape by the AAMC headquarters in Washington. Out of the total group of subjects, foreign students took a different test (Series 48) and were thus excluded from our analyses. The remaining total sample size is therefore made up of 18,075 subjects.

The procedures

Studies of model-data fit concerning the simple logistic model were initiated in America with Wright's paper (1968). The robustness of the model with respect to violations of its assumptions has been well investigated by Panchapakesan (1969) using simulated data. The work of Rentz and Bashaw (1975) on real data has contributed to the clarification of the notion of fit. The basic set of procedures used in this study reflects a great deal of what was suggested by these authors in exploring the fit of the simple logistic model to some test data.

Using the program described in the last section of this chapter, we shall conduct two series of analyses. In the first series, we shall attempt to find out whether or not there is fit between the three chosen MCAT subtests and the simple logistic model. An overall run on each subtest with the total sample shall be used for that purpose. In a stepwise manner, we shall examine the results of the overall tests of fit, chi-square and likelihood ratio, and the results of the item chi-square test of fit. This latter procedure is aimed at identifying the number of conformable items in each subtest and the number of items

responsible for the misfit of the subtest. In looking for the sources of misfit, we shall first examine the size of each score group and identify the number of misfitting items per score group. We shall then proceed further in exploring successively the usual causes of misfit. In the formulation of the model a basic assumption is that the trait being measured is unidimensional. Because of many considerations, the cost involved being of some importance, we shall not factor analyze our data. However, we feel quite confident that the results of such an analysis would show that the MCAT subtests do not violate the unifactor assumption.

> "For tests that appear to be homogeneous like spelling, vocabulary, reading tests, etc., it has been found that when the tetrachoric item intercorrelations were analyzed, in general one factor accounted for most of the item variance." (Panchapakesan, 1969).

The MCAT tests are constructed with care and professional quality. One can then assume that there are a sufficient number of items which satisfy the assumption of item homogeneity. In an analysis of the verbal SAT data, a test similar in nature to the MCAT Verbal subtest, Coffman (1966) found that 66 percent of the item variance was accounted for by the first factor and only 7 percent by the next factor.

The most important assumption of the model is that all items have the same discriminating power. To detect variation in item discrimination, we shall use the procedure described by Panchapakesan (1969). Here, the standardized difference of the estimated slope from unity shall be the criterion of interest to identify a new conformable set of items.

We shall then explore guessing and speed as likely hypotheses of misfit. To achieve this, we shall look at the correlation between

normal deviates and score groups.

In summary, for this first series of analyses which will be presented in Chapter V, the general approach is that suggested by Wright (1969) and the criteria examined are those recommended by Panchapakesan (1969), that is, item chi-square, a slope index of fit (interval 1.0 ± 0.2 and the standardized difference of the slope from unity $|S|$), and the correlation coefficient between normal deviates and score groups.

In Chapter VI, we shall investigate the effects of misfit on test calibration and person measurement. The procedures to be used have been described in Wright's articles (1968, 1969) and partly applied by Rentz and Bashaw (1975). According to the simple logistic model, as explained in Chapter II, the parameters describing the difficulty of the items and the ability of the subjects should be invariant. One way to examine this issue is to divide the overall sample of examinees into various subgroups, to calibrate the test on each subgroup, and to evaluate the magnitude of the differences between easiness and ability estimates. Using such a procedure, we shall conduct two sets of investigations. One type of split was made according to the subjects' responses to question 14 in the AAMC questionnaire:

"14- Estimate and indicate your parents' combined gross annual income for last year.
1. Less than $5,000
2. $5,000 - 9,999
3. $10,000 - 14,999
4. $15,000 - 19,999
5. $20,000 or more"

Table 2 provides the frequencies for each category of income. Categories 3 and 4 were grouped and correspond to category 3 in the table whereas category 5 becomes category 4 in table 2. There are

14,767 usable records.

For the second split, the overall sample was divided according to the subjects' responses to question 16 of the AAMC questionnaire:

"16- How do you describe yourself?
1. Afro-American or Black
2. American Indian or Native American
3. Caucasian or White
4. Mexican-American or Chicano
5. Oriental or Asian-American
6. Puerto Rican (Mainland)
7. Spanish-speaking American
8. Other "

Table 3 provides the frequencies for three subgroups: White (3), Black (1) and Other (2,4,5,6,7,8).

There are 12,599 usable records for this split. In both cases repeaters were not excluded from the analyses.

The choice of the variables for these investigations was made with the intent of constituting samples that differ as much as possible from one another. We shall use a chi-square statistic to test the degree of concordance between log easiness estimates for each item and log ability estimates for each score group. We shall then relate the results of these two sets of analyses with those of Chapter V so that a link is established between indicators of fit at the item level and indicators of fit at the test level.

The log estimates were computed from the specifications of the model, already discussed in Chapter II, using the computer program which will now be briefly described.

TABLE 2

FREQUENCY OF SUBJECTS ACCORDING
TO PARENTS' INCOME LEVEL

INCOME	FREQUENCY
1. Less than $5,000	1,030
2. $5,000 - 9,999	4,620
3. $10,000 - 19,999	6,317
4. $20,000 or more	4,800
Total	14,767

TABLE 3

FREQUENCY OF SUBJECTS ACCORDING
TO RACIAL BACKGROUND

RACE	FREQUENCY
1. White	10,685
2. Black	626
3. Other	1,288
Total	12,599

The computer program

For our investigations, we used the unconditional maximum likelihood item analysis program developed by Wright and Panchapakesan (1969). The original program used the FORTRAN II language and was adapted to operate under the FASTRAN compiler of the University of Chicago IBM 7094/7040 System. The modifications we made consisted in transforming the language to FORTRAN IV, adapting the program to operate under the compiler of Michigan State University CDC 6500 System, and adjusting a few subroutines so that the data could be entered from a tape.

The program capacity is 150 items and 100,000 subjects which is well beyond what we needed. A full description of the program is given in Wright and Panchapakesan (1969). We shall therefore only summarize here the main steps. The program first reads and scores the data from a scoring key provided by the user in response vector format. Then, it generates a simple item by score group count matrix (matrix A) where each cell, a_{ji}, represents the number of people in a given score group j who got item i correct. In this matrix, the row margin contains the total number of subjects who got a right answer to each item and the column margin contains each total score group size. This count matrix is adjusted for non informative items, that is, items either answered correctly or missed by all subjects, and non informative score groups, that is, zero or maximum scores. The easiness and ability estimates are computed from this matrix using the estimation procedure described in Chapter II. A second matrix is then generated by the program. It is a matrix of normal deviates from expectation

(matrix Y) where each cell, y_{ji}, provides the standardized value of the difference between the observed count and the count expected on the basis of the model specifications. The item test of fit described in Chapter II is computed from this matrix. Finally, the program provides a summary which includes for each item the percent of subjects passing, a log easiness estimate (ability intercept of item characteristic curve at median response) and its standard error, a discrimination value (slope of item characteristic curve at median response), a reliability estimate (point biserial correlation between item response and estimated ability), and the probability value associated with the item mean square. It also includes for each score the sample frequency at that score, the sample percentile through that score, a log ability estimate on an interval scale and its standard error, a raw ability estimate on a ratio scale and its confidence boundaries. For the overall test, the program computes a chi-square and a likelihood ratio test.

CHAPTER V

SOURCES OF MISFIT IN THE MCAT TEST

In this chapter, we shall attempt to describe the nature of the misfit between the MCAT subtests and the simple logistic model. To do this, we shall proceed from the general to the specific in a step-wise manner starting with a general overall test of goodness of fit.

The overall test of fit

Each of the three MCAT subtests was first submitted to an overall analysis. According to both tests of fit, the chi-square and likelihood ratio tests, none of the subtests fit the simple logistic response model. These results are not surprising. With a large sample size of 18,075 subjects, almost any deviation from expectation, however small it may be, would be detected by these two statistical tests. Table 4 shows the mean square (as defined in the last section of Chapter II), the probability and the degrees of freedom for the chi-square and likelihood ratio tests applied to each MCAT subtest. The probabilities are all smaller than .001 which reflects a lack of fit between data and model.

After excluding the non informative score groups of zero and all-correct, there were only two empty score groups in the Quantitative Ability subtest ($m = 47$), five in the Verbal Ability subtest ($m = 69$), and eleven in the Science subtest ($m = 74$). There were 1 zero score and 130 maximum scores in the Quantitative Ability subtest for a total of 17,944 good cases left. All cases were good cases for the other two

subtests. Moreover, all of the items were good items, that is, none of them in all three subtests were either answered completely correctly or missed by all of the subjects.

TABLE 4

OVERALL TEST OF FIT BETWEEN DATA AND MODEL

STATISTICS	SUBTEST		
	QUANTITATIVE	VERBAL	SCIENCE
CHI-SQUARE			
Mean square*	5.308	6.156	2.996
Probability	.000	.000	.000
LIKELIHOOD RATIO			
Mean square*	1.129	4.187	9.293
Probability	.000	.000	.000
DEGREES OF FREEDOM	2254	5032	6205
SAMPLE SIZE	18,075	18,075	18,075

*Overall chi-square divided by its degrees of freedom

In a brief presentation made during the 1969 presession on the Rasch model, held in conjunction with the AERA annual meeting, Benjamin Wright suggested that when both tests of fit achieve very low probabilities (below .001), one should find out which items are causing this misfit, delete these items and recalibrate the test.

The problem of fit is not a simple one and Professor Wright suggested three ways of approaching it: 1) One should attempt to find out whether or not parameters estimated from the model can reproduce the original data, 2) whether or not scoring tables based on extremely

different calibration samples are comparable, and 3) whether or not the slopes of the item log odds lines are close to unity.

We shall now explore the first question: do parameters estimated from the model reproduce the data? This is the kind of fit tested by the item chi-square statistic.

The item chi-square test of fit

As Wright states it, the item chi-square test of fit is elegant theoretically but it does not work as well in practice as one might like. The problem is that it is too sensitive, that is, it is significant when in fact the items fit the model well enough for practical purposes. For that reason, Wright recommends not to use a .05 or .01 significance level for item rejection. Another reason for avoiding the use of such a rule is that, because of the large number of items involved, even if there were a good fit between data and model, some items would obtain probabilities less than .05 or .01 due to the random variation which is part of the model. However, Wright suggests that when the probability goes below .001, that is usually a bad sign. Hence, .001 is considered to be a reasonable cut off point.

Table 5 shows the number and percentage of items which would be rejected in each of the three MCAT subtests depending on the statistical criterion chosen. It can be seen that even at the lowest probability level of .001, only a few items would be retained in each subtest.

Because of the sensitivity of the test which is magnified by the size of our sample, one must look for alternative ways of exploring fit.

The 43 items (20%) that fit on the single criterion of $p \geq .001$ for the three subtests are listed in table 6.

Since there is a one to one relationship between the mean squares and the probabilities, one could look at the magnitude of the mean squares to get some clues as to where the misfit is occurring for each item. If the mean square for an item is larger than 1, say 2 or more, it indicates that the estimated parameters are not reproducing the data well for that item.

The computer program produces a table of normal deviates from expectation. These values are a measure of the misfit of a given item for a given score group. According to the model, they have a normal distribution with mean 0 and variance 1. One can then look at the normal deviates and check the score groups for which these values exceed plus or minus 3. As Wright suggests, if the score group is small, less than 10 persons, then a misfit based on that score group is less significant. On the other hand, if a score group is large, 20 persons or more, then a misfit is worth paying attention to.

The distribution of score groups according to their sizes is given in table 7 for the three subtests. Score groups of 0 and maximum scores are excluded. In the three subtests combined, we can see that nearly 75% of the score groups contain more than 20 subjects.

Knowing that as the mean squares approach 1 the probabilities approach .5 and as the mean squares increase towards 2 and 3 the probabilities decrease, it seems reasonable to divide the misfitting items on the basis of the magnitude of their mean squares in order to better pinpoint the source of misfit. This was done for the three subtests and the data are presented in tables 8,9, and 10.

TABLE 5

NUMBER AND PERCENTAGE OF MISFITTING ITEMS
DEPENDING ON THE CRITERION VALUE SELECTED FOR THE ITEM CHI-SQUARE

SUBTEST	PROBABILITY					
	LESS THAN .05		LESS THAN .01		LESS THAN .001	
	N	%	N	%	N	%
Quantitative (K = 50)	45	90	42	84	40	80
Verbal (K = 75)	73	97	72	96	69	92
Science (K = 86)	72	84	71	83	59	69
Total (K = 211)	190	90	185	88	168	80

TABLE 6

LIST OF ITEMS FOR WHICH $p \geq .001$

QUANTITATIVE (N = 10)	VERBAL (N = 6)	SCIENCE (N = 27)		
9	8	1	46	72
15	11	3	47	74
16	17	6	48	76
18	27	20	51	77
20	43	22	53	81
21	73	29	58	85
24		32	61	86
27		39	63	
31		41	66	
42		45	69	

TABLE 7

DISTRIBUTION OF SCORE GROUPS FOR
THE THREE SUBTESTS

SUBTEST	SIZE OF SCORE GROUPS		
	LESS THAN 10	BETWEEN 10 AND 20	20 OR MORE
Quantitative (N = 49)	7	1	41
Verbal (N = 74)	15	2	57
Science (N = 85)	22	7	56
Total (N = 208)	44	10	154 (74%)

The items are rank ordered according to the magnitude of their mean squares, the largest value being first in the list. The probability level is .001 and the score groups are divided into three categories.

The foregoing analysis illustrates the fact that the size of the score groups is not a viable hypothesis to explain the misfit of our data. Only a few items would be retained, were we to disregard their misfit on the basis of score groups of size smaller than 10, which would otherwise be deleted. These items are listed in table 11 along with other items for which the mean square value is less than 2.00. Clearly, these items are not so bad and it will ease the process of exploring the sources of misfit to select them out right away.

From tables 8,9, and 10, it is obvious that the relationship between the magnitude of the mean squares and the number of score groups involved in the misfit is, as expected, very high.

TABLE 8

QUANTITATIVE ABILITY SUBTEST - 40 MISFITTING ITEMS
WITH MEAN SQUARES, NUMBER AND SIZE OF SCORE GROUPS INVOLVED

ITEM	MEAN SQUARE *	SCORE GROUPS			
		LESS THAN 10	BETWEEN 10 AND 20	20 OR MORE	TOTAL
50	31.88	1	0	34	35
22	15.24	0	1	25	26
46	13.51	0	0	24	24
25	12.38	0	0	25	25
43	12.16	1	0	18	19
34	9.06	0	0	19	19
49	7.97	0	0	12	12
10	7.96	0	0	16	16
26	7.91	1	0	16	17
48	6.99	0	0	17	17
1	6.97	0	0	10	10
14	6.91	0	0	14	14
41	6.65	0	1	10	11
30	6.46	1	0	9	10
13	5.92	0	0	14	14
3	5.86	0	0	10	10
6	5.84	0	0	11	11
44	5.75	0	0	11	11
19	5.13	0	0	12	12
12	4.83	1	0	8	9
45	4.69	0	0	10	10
47	4.34	0	1	7	8
11	3.92	0	0	2	2
37	3.90	0	0	5	5
2	3.79	0	0	6	6
36	3.79	1	0	4	5
33	3.79	0	0	7	7
29	3.64	0	0	6	6
39	3.30	0	0	4	4
28	3.05	1	0	3	4
40	3.05	0	0	3	3
4	2.98	0	0	4	4
23	2.74	0	0	1	1
35	2.51	1	0	3	4
17	2.12	0	0	1	1
38	2.09	1	0	0	1
8	1.95	0	0	2	2
32	1.95	0	0	2	2
7	1.90	0	0	2	2
5	1.81	0	0	0	0

* Item chi-square statistic divided by its degrees of freedom

TABLE 9

VERBAL ABILITY SUBTEST - 69 MISFITTING ITEMS WITH MEAN SQUARES, NUMBER AND SIZE OF SCORE GROUPS INVOLVED

ITEM	MEAN SQUARE*	SCORE GROUPS			
		LESS THAN 10	BETWEEN 10 AND 20	20 OR MORE	TOTAL
58	29.08	0	1	41	42
35	21.90	0	0	40	40
75	20.40	1	1	41	43
67	18.02	1	0	35	36
36	16.42	0	0	33	33
4	16.02	0	0	39	39
59	14.04	1	0	35	36
12	10.87	0	0	35	35
65	10.79	0	0	23	23
54	9.90	0	0	24	24
47	9.53	1	0	24	25
49	9.38	0	0	24	24
5	8.53	0	0	24	24
41	8.33	0	0	26	26
71	7.91	0	0	19	19
64	7.84	0	0	28	28
72	7.49	1	0	18	19
6	7.48	0	0	22	22
34	7.30	0	0	23	23
32	6.70	0	1	14	15
44	6.33	3	0	17	20
55	6.26	1	1	13	15
42	6.24	1	0	17	18
30	6.23	2	0	19	21
13	5.93	0	0	17	17
29	5.79	1	0	12	13
48	5.68	0	0	19	19
37	5.61	0	1	18	19
26	5.61	1	0	11	12
3	5.30	0	0	15	15
63	5.21	0	0	13	13
18	5.17	1	0	17	18
57	5.15	0	0	13	13
60	4.81	0	0	15	15
56	4.80	0	1	10	11
31	4.75	0	1	13	14
66	4.71	0	0	11	11
9	4.66	0	0	9	9
53	4.54	1	0	11	12
20	4.54	2	0	8	10

*Item chi-square statistic divided by its degrees of freedom

TABLE 9 - Continued

ITEM	MEAN SQUARE*	SCORE GROUPS			
		LESS THAN 10	BETWEEN 10 AND 20	20 OR MORE	TOTAL
68	4.52	0	0	14	14
51	4.47	0	0	10	10
25	4.33	1	0	10	11
15	4.30	0	1	12	13
14	4.29	1	0	9	10
50	4.06	1	0	7	8
7	4.04	0	0	9	9
16	4.03	2	0	7	9
22	3.99	1	1	6	8
21	3.85	0	0	5	5
52	3.64	0	0	10	10
40	3.53	0	0	10	10
39	3.52	3	0	6	9
28	3.37	1	0	6	7
23	3.30	0	0	8	8
70	3.08	0	0	6	6
24	2.98	1	0	5	6
61	2.75	0	0	6	6
1	2.72	0	0	6	6
10	2.62	1	0	4	5
74	2.33	0	0	4	4
45	2.28	0	0	4	4
19	2.27	0	1	4	5
33	2.24	0	0	1	1
38	2.13	0	0	3	3
69	2.03	0	0	3	3
2	2.00	0	0	1	1
46	1.74	0	0	2	2
62	1.72	0	0	3	3

* Item chi-square statistic divided by its degrees of freedom

TABLE 10

SCIENCE SUBTEST - 59 MISFITTING ITEMS WITH MEAN SQUARES, NUMBER AND SIZE OF SCORE GROUPS INVOLVED

ITEM	MEAN SQUARE*	SCORE GROUPS			
		LESS THAN 10	BETWEEN 10 AND 20	20 OR MORE	TOTAL
15	12.67	1	1	37	39
55	10.49	2	2	31	35
59	9.54	0	0	29	29
84	7.68	2	0	23	25
57	7.07	1	1	23	25
14	6.66	0	0	23	23
17	6.41	0	0	24	24
21	6.06	0	0	18	18
23	5.67	1	0	14	15
9	5.58	1	0	17	18
34	4.79	0	0	12	12
35	4.74	1	1	12	14
10	4.73	0	0	14	14
56	4.70	0	0	13	13
7	4.64	2	0	15	17
62	4.59	0	0	15	15
26	4.15	0	0	8	8
67	3.92	0	0	12	12
64	3.87	0	1	9	10
12	3.70	0	0	6	6
70	3.56	0	1	10	11
49	3.54	2	0	8	10
27	3.52	0	4	4	8
16	3.46	0	0	9	9
18	3.40	0	0	7	7
79	3.29	0	1	7	8
13	3.23	0	1	7	8
24	3.15	0	0	7	7
78	3.07	0	0	6	6
44	3.00	1	0	5	6
38	2.98	0	0	5	5
33	2.96	0	1	5	6
5	2.93	0	1	7	8
50	2.86	0	0	5	5
73	2.85	0	0	4	4
82	2.73	0	0	6	6
80	2.69	1	0	2	3
2	2.52	0	0	2	2
40	2.51	0	0	6	6
25	2.49	0	0	4	4

* Item chi-square statistic divided by its degrees of freedom

TABLE 10 - Continued

ITEM	MEAN SQUARE*	SCORE GROUPS			
		LESS THAN 10	BETWEEN 10 AND 20	20 OR MORE	TOTAL
31	2.36	0	0	2	2
43	2.31	1	1	3	5
75	2.28	0	0	3	3
11	2.27	0	0	1	1
83	2.17	0	0	2	2
37	2.15	0	0	2	2
36	2.13	1	0	2	3
68	2.12	1	1	0	2
71	2.12	0	1	1	2
65	2.11	0	0	1	1
42	2.00	1	0	2	3
54	1.97	1	0	0	1
60	1.95	0	0	1	1
19	1.92	0	1	2	3
8	1.83	1	0	0	1
4	1.80	0	0	2	2
30	1.65	1	0	1	2
28	1.62	1	0	2	3
52	1.61	2	0	1	3

* Item chi-square statistic divided by its degrees of freedom

TABLE 11

LIST OF ITEMS SHOWING VERY SLIGHT SIGNS OF MISFIT

QUANTITATIVE (N = 5)	VERBAL (N = 2)	SCIENCE (N = 8)
5	46	4
7	62	8
8		19
32		28
38		30
		52
		54
		60

Before concluding this section, tables 12,13, and 14 show the actual size of each score group and the number of misfitting items for each one of them. It is worth noting that, in general, the number of misfitting items per score group is quite small. We shall look again at these figures when we examine the ability estimates.

We shall now explore the range of item discriminations in each of the three MCAT subtests.

Item discrimination

We consider a second question: are the slopes of the item log odds lines close to unity? In Chapter II, the form of the simple logistic model and the general logistic model in which items can differ in their discriminations was given. The general logistic model cannot be solved without making further assumptions regarding either the ability

TABLE 12

QUANTITATIVE ABILITY SUBTEST - SIZE OF SCORE GROUPS AND NUMBER OF MISFITTING ITEMS PER SCORE GROUP

SCORE GROUP NO	SIZE	MISFITTING ITEMS	SCORE GROUP NO	SIZE	MISFITTING ITEMS	SCORE GROUP NO	SIZE	MISFITTING ITEMS
1	0	0	18	155	9	35	741	11
2	0	0	19	213	13	36	810	9
3	2	3	20	231	9	37	795	6
4	2	0	21	264	12	38	769	5
5	2	3	22	321	12	39	749	14
6	4	1	23	323	12	40	709	8
7	8	3	24	368	8	41	676	9
8	15	4	25	394	5	42	721	18
9	21	3	26	473	7	43	645	15
10	35	7	27	506	4	44	614	14
11	37	5	28	527	5	45	545	18
12	55	11	29	583	4	46	522	15
13	80	12	30	634	6	47	420	15
14	91	12	31	714	3	48	328	8
15	125	14	32	706	7	49	269	5
16	113	11	33	719	5			
17	176	14	34	734	7			

TABLE 13

VERBAL ABILITY SUBTEST - SIZE OF SCORE GROUPS AND NUMBER OF MISFITTING ITEMS PER SCORE GROUP

SCORE GROUP NO	SCORE GROUP SIZE	MISFITTING ITEMS	SCORE GROUP NO	SCORE GROUP SIZE	MISFITTING ITEMS	SCORE GROUP NO	SCORE GROUP SIZE	MISFITTING ITEMS
1	0	0	27	243	31	51	448	16
2	0	0	28	276	20	52	449	21
3	0	0	29	311	25	53	398	16
4	0	0	30	331	25	54	415	20
5	0	0	31	343	20	55	409	26
6	1	1	32	395	21	56	380	19
7	1	3	33	421	17	57	360	24
8	2	5	34	425	18	58	320	25
9	4	2	35	426	19	59	325	27
10	2	6	36	495	13	60	307	24
11	3	4	37	532	5	61	289	32
12	4	3	38	526	10	62	246	19
13	5	0	39	516	8	63	251	23
14	7	5	40	529	10	64	208	20
15	13	7	41	546	9	65	190	21
16	28	11	42	548	11	66	159	22
17	29	15	43	548	11	67	155	16
18	48	13	44	582	14	68	106	11
19	53	22	45	550	13	69	95	13
20	64	24	46	571	17	70	82	9
21	74	19	47	542	16	71	66	8
22	107	24	48	531	11	72	25	3
23	122	26	49	527	12	73	16	3
24	168	31	50	509	15	74	9	1
25	179	22						
26	230	31						

TABLE 14

SCIENCE SUBTEST - SIZE OF SCORE GROUPS AND NUMBER OF MISFITTING ITEMS PER SCORE GROUP

SCORE GROUP NO	SCORE GROUP SIZE	MISFITTING ITEMS	SCORE GROUP NO	SCORE GROUP SIZE	MISFITTING ITEMS	SCORE GROUP NO	SCORE GROUP SIZE	MISFITTING ITEMS
1	0	0	30	56	3	59	622	9
2	0	0	31	92	11	60	596	10
3	0	0	32	127	10	61	538	9
4	0	0	33	116	11	62	555	10
5	0	0	34	128	5	63	510	15
6	0	0	35	186	14	64	528	19
7	1	4	36	184	19	65	418	18
8	0	0	37	208	13	66	399	17
9	0	0	38	214	10	67	375	19
10	0	0	39	239	10	68	351	17
11	0	0	40	268	11	69	317	20
12	1	3	41	337	13	70	267	19
13	1	4	42	363	14	71	237	18
14	3	0	43	365	9	72	214	9
15	1	1	44	413	12	73	151	8
16	3	2	45	425	6	74	101	7
17	2	2	46	477	4	75	108	10
18	10	4	47	497	3	76	55	7
19	9	2	48	499	2	77	55	4
20	13	4	49	550	3	78	34	3
21	13	2	50	614	7	79	36	1
22	10	1	51	600	2	80	18	2
23	17	1	52	584	6	81	12	3
24	29	29	53	638	0	82	9	3
25	30	30	54	639	0	83	4	2
26	30	30	55	557	1	84	5	2
27	44	44	56	612	1	85	0	0
28	58	58	57	637	2			
29	73	73	58	587	6			

distribution of the sample or the homogeneity of item discrimination. In the Rasch model the latter assumption is preferred because there is then no need to worry about getting a proper standardizing sample and the estimation procedure is much more practical.

To detect variation in item discrimination, we shall use the method described by Panchapakesan (1969). Our purpose is to verify whether or not her results on simulated data could be generalized to real data and whether or not her criteria can effectively help identify discrepant items.

In the general logistic model,

$$P_{ni} = \exp(c_i b_n + d_i) \Big/ (1 + \exp(c_i b_n + d_i))$$

where c_i = item discrimination

d_i = item easiness

b_n = ability of subject

The odds for success are given by

$$O_{ni} = \exp(c_i b_n + d_i)$$

and the log odds by

$$\log O_{ni} = c_i b_n + d_i \qquad (5.1)$$

In order to estimate $\log O_{ni}$, we group together subjects with the same score r and consider them to represent repeated observations of subjects with the same average ability b_n for the score level r.

Then, (5.1) can be written as

$$\log O_{ri} = c_i b_n + d_i \qquad (5.2)$$

An estimate of $\log O_{ri}$ is given by t_{ri}

where $t_{ri} = \log \left(s_{ri} \Big/ (N_r - s_{ri})\right)$ (5.3)

and N_r = the number of persons in score group r,

s_{ri} = the number of persons in score group r who get item i correct

Therefore,

$$t_{ri} = c_i^* b_n^* + d_i^* \simeq c_i b_n + d_i \tag{5.4}$$

Summing over i we get

$$t_{r.} = c_.^* b_n^* + d_.^* \tag{5.5}$$

where $c_.^* = (1/k) \sum_i^k c_i^*$ for k items

and $d_.^* = (1/k) \sum_i^k d_i^*$

Solving (5.5) for b_n^* and substituting in (5.4) we get

$$t_{ri} = (c_i^* \; t_{r.} \Big/ c_.^*) - (c_i^* \; d_.^* \Big/ c_.^*) + d_i^* \tag{5.6}$$

After normalizing $d_.^* = 0$ and $c_.^* = 1$, (5.6) becomes

$$t_{ri} = c_i^* t_{r.} + d_i^* \tag{5.7}$$

If we regress the observed log odds t_{ri} on their score means $t_{r.}$ over the scores r = 1, k-1, the slope, c_i^*, is an estimate of item discrimination.

These values are reported by the computer program used for our analyses along with their standard error of estimation for all items.

Since the magnitude of the standardized difference of the estimated slope from unity should be large for items with discriminations significantly different from the average discrimination of the rest of the items, this information will be used with the item chi-square to detect variation in discrimination in our data as it was for Panchapakesan's data.

The standardized difference of the estimated slope from unity, S, is given by $S_i = (c_i^* - 1) \Big/ S.E._{c_i^*}$

where c_i^* = the estimated slope

$S.E._{c_i^*}$ = the standard error of estimation of the slope

We shall now turn our attention to our data. Table 15 shows the actual distribution of item discriminations for the three MCAT subtests.

TABLE 15

DISTRIBUTION OF ITEM DISCRIMINATIONS IN THE THREE MCAT SUBTESTS

SUBTEST	RANGE OF DISCRIMINATION (c_i^*)							
	$0<c_i<0.4$ $c_i > 2.50$		$0.4\leq c_i<0.6$ $1.66\leq c_i\leq 2.50$		$0.6\leq c_i<0.8$ $1.25<c_i\leq 1.66$		$0.8\leq c_i<1.0$ $1.0\leq c_i\leq 1.25$	
	N	%	N	%	N	%	N	%
Quantitative (K = 50)	2	4	9	18	24	48	15	30
Verbal (K = 75)	10	13	19	25	23	31	23	30
Science (K = 86)	10	12	26	30	24	28	26	30
Total (K = 211)	22	10	54	26	71	34	64	30

This is the kind of distribution which should be expected from empirical studies conducted on other tests of similar nature and quality. The maximum variation in discrimination found in actual test data is in the range of 0.4 to 2.5 (Ross, 1966; Lord, 1968). Only 22 items or 10% of our pooled total of items are outside that range. The minimum and maximum values are, for the Quantitative subtest 0.329 and 1.176, for the Verbal subtest 0.133 and 1.538, and for the Science subtest 0.114 and 1.317.

We shall now look at how effective the two selected criteria, item chi-square and standardized difference of the slope from unity, are in identifying discrepant items. For the standardized difference of the slope from unity (S), Panchapakesan suggests taking the absolute value of 3 as the highest acceptable limit for fit. The number of fitting and misfitting items on the basis of both criteria for the three subtests combined is given in table 16.

Our data are, in general, consistent with Panchapakesan's findings. All the items in the range 0-0.4 show misfit on both counts. Only the items in the range of 0.8 to 1.25 show fit on both counts. In the middle range of 0.4 to 0.6, most items present a double misfit.

Table 17 provides the data for the subset of items in the range 0-0.4 for the three subtests. This group of items would therefore be the worst in the three MCAT subtests. The data for the other ranges are presented in tables 18 to 26.

However, one may wonder about the choice of $|3|$ as a cut off point for the standardized difference of the estimated slope from unity. In Panchapakesan's simulations, this value seems to have been arrived at on the basis of a visual approach to fit provided by two kinds of plots,

TABLE 16

NUMBER OF FITTING AND MISFITTING ITEMS FOR
THE THREE SUBTESTS COMBINED

RANGE OF DISCRIMINATIONS	DOUBLE FIT $p \geq .001$ $\lvert S \rvert \leq 3$	SINGLE FIT		DOUBLE MISFIT $p < .001$ $\lvert S \rvert > 3$
		$p < .001$ $\lvert S \rvert \leq 3$	$p \geq .001$ $\lvert S \rvert > 3$	
0.0 - 0.4 (N = 22)	0	0	0	22 (100%)
0.4 - 0.6 (N = 54)	0	0	9	45 (83%)
0.6 - 0.8 (N = 71)	0	0	25	46 (71%)
0.8 - 1.0 (N = 64)	7	45	2	10 (16%)
TOTAL (N = 211)	7 (3%)	45 (21%)	36 (17%)	123 (58%)

TABLE 17

ITEM CHI-SQUARE PROBABILITY AND STANDARDIZED DIFFERENCE OF SLOPE FROM UNITY FOR ITEM DISCRIMINATIONS IN THE RANGE 0-0.4

ITEM		SLOPE	ERROR	|S|	p
QUANTITATIVE ABILITY					
(N = 2)	22	0.329	0.032	20.96	0.000
	50	0.335	0.039	17.05	0.000
VERBAL ABILITY	58	0.133	0.024	36.12	0.000
(N = 10)	75	0.163	0.028	29.89	0.000
	67	0.177	0.027	30.48	0.000
	71	0.304	0.036	19.33	0.000
	65	0.338	0.030	22.06	0.000
	59	0.356	0.026	24.76	0.000
	48	0.364	0.040	15.90	0.000
	72	0.366	0.037	17.13	0.000
	55	0.378	0.035	17.77	0.000
	53	0.389	0.044	13.88	0.000
SCIENCE	55	0.114	0.028	31.64	0.000
(N = 10)	84	0.258	0.039	19.02	0.000
	15	0.313	0.044	15.61	0.000
	14	0.327	0.045	14.95	0.000
	35	0.341	0.033	19.96	0.000
	80	0.359	0.036	17.80	0.000
	49	0.369	0.042	15.02	0.000
	57	0.374	0.042	14.90	0.000
	64	0.385	0.043	14.30	0.000
	43	0.397	0.046	13.40	0.000

TABLE 18

QUANTITATIVE ABILITY SUBTEST - ITEM CHI-SQUARE PROBABILITY AND STANDARDIZED DIFFERENCE OF SLOPE FROM UNITY FOR ITEM DISCRIMINATIONS IN THE RANGE 0.4-0.6

ITEM (N = 9)	SLOPE	ERROR	|S|	p	MISFIT WITH		
					p	|S|	p and |S|
1	0.561	0.034	12.91	0.000	*	*	*
12	0.561	0.040	10.97	0.000	*	*	*
14	0.488	0.032	16.00	0.000	*	*	*
28	0.512	0.050	9.76	0.000	*	*	*
26	0.459	0.050	10.82	0.000	*	*	*
35	0.532	0.044	10.63	0.000	*	*	*
30	0.469	0.048	11.06	0.000	*	*	*
41	0.531	0.045	10.42	0.000	*	*	*
46	0.461	0.037	14.56	0.000	*	*	*
			TOTAL		9	9	9

TABLE 19

VERBAL ABILITY SUBTEST - ITEM CHI-SQUARE PROBABILITY AND STANDARDIZED DIFFERENCE OF SLOPE FROM UNITY FOR ITEM DISCRIMINATIONS IN THE RANGE 0.4-0.6

ITEM (N=19)	SLOPE	ERROR	|S|	p	MISFIT WITH		
					p	|S|	p and |S|
60	0.549	0.038	11.86	0.000	*	*	*
64	0.526	0.042	11.28	0.000	*	*	*
42	0.429	0.040	14.27	0.000	*	*	*
40	0.554	0.045	9.91	0.000	*	*	*
10	0.443	0.043	12.95	0.000	*	*	*
46	0.549	0.041	11.00	0.000	*	*	*
66	0.419	0.044	13.20	0.000	*	*	*
45	0.487	0.043	11.93	0.000	*	*	*
44	0.409	0.044	13.43	0.000	*	*	*
24	0.489	0.048	10.64	0.000	*	*	*
68	0.517	0.052	9.28	0.000	*	*	*
28	0.557	0.045	9.84	0.000	*	*	*
50	0.580	0.049	8.57	0.000	*	*	*
74	0.557	0.041	10.80	0.000	*	*	*
73	0.591	0.034	12.02	0.001		*	
22	0.451	0.052	10.55	0.000	*	*	*
54	0.491	0.051	9.98	0.000	*	*	*
47	0.545	0.049	9.28	0.000	*	*	*
29	0.537	0.046	10.06	0.000	*	*	*
			TOTAL		18	19	18

TABLE 20

SCIENCE SUBTEST - ITEM CHI-SQUARE PROBABILITY AND STANDARDIZED DIFFERENCE OF SLOPE FROM UNITY FOR ITEM DISCRIMINATIONS IN THE RANGE 0.4-0.6

ITEM (N = 26)	SLOPE	ERROR	|S|	p	MISFIT WITH		
					p	|S|	p and |S|
56	0.487	0.040	12.82	0.000	*	*	*
16	0.498	0.039	12.87	0.000	*	*	*
27	0.438	0.034	16.52	0.000	*	*	*
29	0.571	0.043	9.97	0.004		*	
28	0.535	0.038	12.23	0.000	*	*	*
78	0.486	0.036	14.27	0.000	*	*	*
79	0.552	0.045	9.95	0.000	*	*	*
5	0.542	0.043	10.65	0.000	*	*	*
7	0.446	0.040	13.85	0.000	*	*	*
33	0.557	0.044	9.84	0.000	*	*	*
37	0.556	0.051	8.70	0.000	*	*	*
45	0.575	0.048	8.85	0.032		*	
51	0.575	0.050	8.50	0.099		*	
42	0.426	0.039	14.71	0.000	*	*	*
47	0.544	0.049	9.30	0.112		*	
36	0.437	0.041	13.73	0.000	*	*	*
13	0.416	0.043	13.58	0.000	*	*	*
46	0.453	0.051	10.72	0.001		*	
52	0.521	0.051	9.39	0.000	*	*	*
63	0.515	0.050	9.70	0.002		*	
58	0.456	0.043	12.65	0.004		*	
60	0.447	0.050	11.06	0.000	*	*	*
86	0.531	0.045	10.42	0.213		*	
83	0.468	0.046	11.56	0.000	*	*	*
68	0.574	0.049	8.69	0.000	*	*	*
70	0.450	0.042	13.09	0.000	*	*	*
			TOTAL		18	26	18

TABLE 21

QUANTITATIVE ABILITY SUBTEST - ITEM CHI-SQUARE PROBABILITY AND STANDARDIZED DIFFERENCE OF SLOPE FROM UNITY FOR ITEM DISCRIMINATIONS IN THE RANGE 0.6-0.8

ITEM (N = 24)	SLOPE	ERROR	|S|	p	MISFIT WITH		
					p	|S|	p and |S|
2	0.789	0.056	3.76	0.000	*	*	*
3	0.709	0.041	7.09	0.000	*	*	*
8	0.799	0.043	4.67	0.000	*	*	*
16	0.696	0.040	7.60	0.077		*	
4	0.644	0.041	8.68	0.000	*	*	*
15	0.622	0.045	8.40	0.001		*	
36	0.717	0.050	5.66	0.000	*	*	*
27	0.774	0.053	4.26	0.061		*	
20	0.650	0.034	10.29	0.001		*	
33	0.786	0.043	4.97	0.000	*	*	*
21	0.714	0.050	5.72	0.062		*	
24	0.754	0.038	6.47	0.031		*	
31	0.704	0.046	6.43	0.013		*	
40	0.787	0.041	5.19	0.000	*	*	*
38	0.660	0.054	6.29	0.000	*	*	*
37	0.769	0.045	5.13	0.000	*	*	*
29	0.754	0.055	4.47	0.000	*	*	*
42	0.747	0.033	7.66	0.394		*	
45	0.764	0.051	4.62	0.000	*	*	*
47	0.650	0.045	7.77	0.000	*	*	*
32	0.787	0.044	4.84	0.000	*	*	*
39	0.626	0.049	7.63	0.000	*	*	*
48	0.749	0.049	5.12	0.000	*	*	*
49	0.603	0.044	9.02	0.000	*	*	*
			TOTAL		16	24	16

TABLE 22

VERBAL ABILITY SUBTEST - ITEM CHI-SQUARE PROBABILITY AND STANDARDIZED DIFFERENCE OF SLOPE FROM UNITY FOR ITEM DISCRIMINATIONS IN THE RANGE 0.6 - 0.8 AND 1.25 - 1.66

ITEM (N = 23)	SLOPE	ERROR	|S|	p	MISFIT WITH		
					p	|S|	p and |S|
32	1.370	0.074	5.00	0.000	*	*	*
56	1.280	0.055	5.09	0.000	*	*	*
37	0.698	0.048	6.29	0.000	*	*	*
36	1.448	0.096	4.66	0.000	*	*	*
5	1.405	0.077	5.25	0.000	*	*	*
4	1.538	0.099	5.43	0.000	*	*	*
38	0.783	0.047	4.61	0.000	*	*	*
11	0.684	0.054	5.85	0.010		*	
35	1.524	0.061	8.59	0.000	*	*	*
62	0.610	0.050	7.80	0.000	*	*	*
39	0.713	0.058	4.94	0.000	*	*	*
43	0.763	0.045	5.26	0.539		*	
17	0.653	0.034	10.20	0.001		*	
15	0.742	0.048	5.37	0.000	*	*	*
16	0.660	0.059	5.76	0.000	*	*	*
19	0.712	0.045	6.40	0.000	*	*	*
20	0.621	0.056	6.76	0.000	*	*	*
21	0.684	0.056	5.64	0.000	*	*	*
70	0.624	0.044	8.54	0.000	*	*	*
26	0.708	0.049	5.95	0.000	*	*	*
51	0.705	0.053	5.56	0.000	*	*	*
27	0.624	0.053	7.09	0.001		*	
30	0.633	0.060	6.11	0.000	*	*	*
			TOTAL		19	23	19

TABLE 23

SCIENCE SUBTEST - ITEM CHI-SQUARE PROBABILITY AND STANDARDIZED DIFFERENCE OF SLOPE FROM UNITY FOR ITEM DISCRIMINATIONS IN THE RANGE 0.6 - 0.8 AND 1.25 - 1.66

ITEM (N = 24)	SLOPE	ERROR	|S|	p	MISFIT WITH		
					p	|S|	p and |S|
21	1.317	0.066	4.80	0.000	*	*	*
8	0.672	0.053	6.19	0.000	*	*	*
4	0.686	0.045	7.00	0.000	*	*	*
77	0.659	0.046	7.57	0.001		*	
30	0.797	0.064	3.17	0.000	*	*	*
32	0.667	0.050	6.66	0.009		*	
85	0.763	0.050	4.74	0.222		*	
39	0.649	0.046	7.80	0.055		*	
41	0.754	0.054	4.55	0.355		*	
76	0.730	0.047	5.74	0.737		*	
44	0.774	0.059	3.83	0.000	*	*	*
53	0.642	0.049	7.30	0.406		*	
48	0.656	0.046	7.64	0.468		*	
54	0.730	0.041	6.58	0.000	*	*	*
61	0.738	0.050	5.24	0.001		*	
66	0.678	0.046	7.00	0.311		*	
18	0.622	0.052	7.26	0.000	*	*	*
69	0.721	0.040	6.80	0.542		*	
65	0.710	0.044	6.59	0.000	*	*	*
81	0.613	0.048	8.06	0.003		*	
73	0.642	0.047	7.61	0.000	*	*	*
67	0.746	0.056	4.53	0.000	*	*	*
71	0.734	0.055	4.83	0.000	*	*	*
72	0.737	0.060	4.38	0.001		*	
			TOTAL		11	24	11

TABLE 24

QUANTITATIVE ABILITY SUBTEST - ITEM CHI-SQUARE PROBABILITY AND STANDARDIZED DIFFERENCE OF SLOPE FROM UNITY FOR ITEM DISCRIMINATIONS IN THE RANGE 0.8 - 1.25

ITEM (N = 15)	SLOPE	ERROR	|S|	p	MISFIT WITH		
					p	|S|	p and |S|
5	0.925	0.046	1.63	0.000	*		
7	0.954	0.044	1.04	0.000	*		
9	0.872	0.047	2.72	0.024			
6	1.038	0.046	0.82	0.000	*		
10	1.168	0.069	2.43	0.000	*		
11	0.978	0.052	0.42	0.000	*		
23	0.824	0.048	3.66	0.000	*	*	*
13	1.046	0.035	1.31	0.000	*		
18	0.820	0.036	5.00	0.686		*	
34	1.176	0.048	3.66	0.000	*	*	*
19	0.940	0.040	1.50	0.000	*		
25	1.031	0.051	0.60	0.000	*		
44	0.891	0.057	1.91	0.000	*		
17	0.840	0.056	2.85	0.000	*		
43	0.809	0.060	3.18	0.000	*	*	*
			TOTAL		13	4	3

TABLE 25

VERBAL ABILITY SUBTEST - ITEM CHI-SQUARE PROBABILITY AND STANDARDIZED DIFFERENCE OF SLOPE FROM UNITY FOR ITEM DISCRIMINATIONS IN THE RANGE 0.8 - 1.25

ITEM (N = 23)	SLOPE	ERROR	|S|	p	MISFIT WITH		
					p	|S|	p and |S|
1	1.177	0.099	1.78	0.000	*		
31	0.981	0.075	0.25	0.000	*		
33	0.912	0.052	1.69	0.000	*		
2	0.915	0.045	1.88	0.000	*		
3	1.169	0.054	3.12	0.000	*	*	*
57	1.065	0.047	1.38	0.000	*		
7	1.008	0.072	0.11	0.000	*		
6	1.185	0.067	2.76	0.000	*		
9	0.952	0.073	0.65	0.000	*		
34	1.023	0.055	0.41	0.000	*		
8	0.827	0.053	3.26	0.195		*	
63	1.002	0.049	0.04	0.000	*		
61	0.989	0.052	0.21	0.000	*		
12	1.235	0.074	3.17	0.000	*	*	*
13	1.021	0.066	0.31	0.000	*		
41	0.948	0.075	0.69	0.000	*		
69	0.908	0.046	2.00	0.000	*		
18	0.830	0.048	3.54	0.000	*	*	*
14	0.944	0.050	1.12	0.000	*		
23	0.801	0.057	3.49	0.000	*	*	*
25	0.902	0.058	1.68	0.000	*		
49	0.930	0.064	1.09	0.000	*		
52	0.903	0.053	1.83	0.000	*		
			TOTAL		22	5	4

TABLE 26

SCIENCE SUBTEST - ITEM CHI-SQUARE PROBABILITY AND STANDARDIZED DIFFERENCE OF SLOPE FROM UNITY FOR ITEM DISCRIMINATIONS IN THE RANGE 0.8 - 1.25

ITEM	SLOPE	ERROR	|S|	p	MISFIT WITH		
					p	|S|	p and |S|
6	1.001	0.071	0.01	0.006			
11	1.100	0.067	1.49	0.000	*		
1	0.849	0.070	2.16	0.001			
3	0.950	0.041	1.22	0.833			
12	1.109	0.051	2.14	0.000	*		
20	0.930	0.051	1.37	0.164			
19	0.968	0.055	0.58	0.000	*		
2	1.220	0.057	3.85	0.000	*	*	*
26	1.029	0.068	0.43	0.000	*		
24	0.829	0.044	3.88	0.000	*	*	*
75	0.844	0.052	3.00	0.000	*		
40	1.002	0.063	0.03	0.000	*		
74	0.948	0.046	1.15	0.576			
22	0.861	0.051	2.72	0.002			
17	0.918	0.066	1.24	0.000	*		
34	1.144	0.056	2.57	0.000	*		
25	1.035	0.049	0.71	0.000	*		
9	1.090	0.062	1.45	0.000	*		
23	0.905	0.065	1.46	0.000	*		
38	0.945	0.061	0.90	0.000	*		
10	1.037	0.049	0.75	0.000	*		
31	0.906	0.058	1.62	0.000	*		
82	0.816	0.054	3.40	0.000	*	*	*
50	0.939	0.064	0.95	0.000	*		
59	0.916	0.069	1.21	0.000	*		
62	1.008	0.054	0.15	0.000	*		
			TOTAL		20	3	3

a probability plot where the observed standardized difference was plotted against the normal deviate corresponding to the proportion (2i-1)/2n (i is the rank of the value of the observation and n is the total number of observations), and a gamma plot where the observed mean square for each item was plotted against the expected mean square.

Panchapakesan comments that probability plotting is a subjective method in that the determination of whether or not the data fit is based on a visual examination rather than a statistical calculation.

We believe that there is a statistical rationale which can be provided to explain why the simple logistic model treats items with discriminations in the range 0.8-1.25 as homogeneous. That rationale could also explain why some of the items outside that range are not always depicted as poor fitting by the item chi-square test of fit whereas some of the items inside that range could yield values of $|S|$ greater than 3.

In her work, Panchapakesan does not always report the value of the standard error of the slope. A quick check however shows that it is, in general, quite small. This value is crucial since it affects directly the magnitude of $|S|$. It is only in an error-free context that one could establish a limiting value of the slope in absolute terms of say 0.8. For instance, if the slope is estimated as 0.900 and the standard error is 0.02, the value of $|S|$ is then 5.00, thus exceeding 3, and the item will be classified in the misfit category. If, on the other hand, for an estimated slope of 0.900 the standard error is 0.04, $|S|$ will then be 2.50 and the item will be classified as fitting the model. At the other extreme, a slope of 0.400 will satisfy the cut off value of $|S| = 3$ if its standard error is equal to or exceeds 0.20.

This issue is important for item selection. It indicates that items may fit even if their discriminations look quite discrepant from unity.

Since the range 0.8-1.2 was consistently the best range in Panchapakesan's studies and since the value of 3 for |S| seemed a good criterion for detecting discrepant items, we can derive the following statistical conclusion: all items for which the estimated slope falls within plus or minus two standard errors around unity are treated as homogeneous items by the simple logistic model.

To construct a 95 percent confidence interval around one is statistically sensible. Then the values 0.8 and 1.2 correspond to two standard errors around the estimated value 1.0 when the magnitude of the standard error is 0.10.

If we accept that a standard error of 0.1 or less is close enough to an error-free context, we can explain why it has been possible for Panchapakesan to establish an absolute value of 0.8 as a limiting criterion for an acceptable variation in item discriminations.

In our data, the standard errors are all very small. The range is .024-.099 for the 211 items. Table 27 shows the minimum value of the standard error corresponding to different ranges of discriminations for |S| to be equal to or smaller than 3. This table could be expanded to include negative slopes. We can see that for all values of the slope smaller than 0.700, the standard error needs to exceed 0.1 for |S| to be equal to or smaller than 3.

For practical applications, we conclude from the discussion above that one must first consider the value of .8 - 1.2 as the criterion of

TABLE 27

EXPECTED VALUE OF THE STANDARD ERROR OF THE ESTIMATED SLOPE FOR A STANDARDIZED DIFFERENCE OF SLOPE FROM UNITY EQUAL TO OR SMALLER THAN |3|

RANGE OF DISCRIMINATIONS	STANDARD ERROR $\geq$ THAN	RANGE OF DISCRIMINATIONS	STANDARD ERROR $\geq$ THAN
.999-.970	0.01	.489 - .460	0.18
.969-.940	0.02	.459 - .430	0.19
.939-.910	0.03	.429 - .400	0.20
.909-.880	0.04	.399 - .370	0.21
.879-.850	0.05	.369 - .340	0.22
.849-.820	0.06	.339 - .310	0.23
.819-.790	0.07	.309 - .280	0.24
.789-.760	0.08	.279 - .250	0.25
.759-.730	0.09	.249 - .220	0.26
.729-.700	0.10	.219 - .190	0.27
.699-.670	0.11	.189 - .160	0.28
.669-.640	0.12	.159 - .130	0.29
.639-.610	0.13	.129 - .100	0.30
.609-.580	0.14	.099 - .070	0.31
.579-.550	0.15	.069 - .040	0.32
.549-.520	0.16	.039 - .010	0.33
.519-.490	0.17	negative values	0.34 and +

fit since $|S| > 3$ is too stringent a rule when the standard errors are very small.

To obtain more compatibility with Panchapakesan's findings and better demonstrate the soundness of her criteria, tables 18 to 26 were reanalyzed with an adjusted $|S|$. For each value of the slope equal to or greater than 0.700, the standard error was set at a fixed value of 0.1 and $|S|$ recomputed. Table 28 provides the summary data. Of course, the distribution remains unchanged for the range 0.0-0.6. The number of items with double misfit drops from 71 percent to 27 percent in the range 0.6-0.8 and from 16 percent to zero in the range 0.8-1.0. And this is what should be expected in a probabilistic sense. A 95 percent

TABLE 28

NUMBER OF FITTING AND MISFITTING ITEMS FOR
THE THREE SUBTESTS COMBINED WITH AN ADJUSTED |S|

RANGE OF DISCRIMINATIONS	DOUBLE FIT $p \geq .001$ $\lvert S \rvert^* \leq 3$	SINGLE FIT		DOUBLE MISFIT $p < .001$ $\lvert S \rvert^* > 3$
		$p < .001$ $\lvert S \rvert^* \leq 3$	$p \geq .001$ $\lvert S \rvert^* > 3$	
0.0 - 0.4 (N = 22)	0	0	0	22 (100%)
0.4 - 0.6 (N = 54)	0	0	9	45 (83%)
0.6 - 0.8 (N = 71)	12	27	13	19 (27%)
0.8 - 1.0 (N = 64)	9	55	0	0 (0%)
TOTAL (N = 211)	21 (10%)	82 (39%)	22 (10%)	86 (41%)

Key: $\lvert S \rvert^*$ = adjusted $\lvert S \rvert$ for values of slope $\bar{>}$ 0.700 for which S.E. $\bar{<}$ 0.10

confidence interval around 0.8 goes from 0.6 to 1.0. But even a 99 percent confidence interval around 0.4 will never reach 0.8 (6.58) when the standard error is 0.1. We conclude that the procedure derived by Panchapakesan on simulated data is sound and can be applied with confidence to real data. We would only suggest to be careful not to reject items in the range of discriminations 0.7-1.0 when the standard error of estimation is smaller than 0.1 if |S| is used as a criterion of fit. This conclusion is only tentative at this point. We need to assess the effect of such a variation in item discriminations on calibration and measurement which is the subject of the next chapter.

We started off with 80% of the items showing a lack of fit (Table 5). After looking at the mean squares and the size of the score groups involved, that percentage was lowered to 73% (Table 11). We then introduced a new statistic |S|on the basis of which the misfitting items were reduced to 58% (Table 16). Finally, adjusting |S| for small standard errors, the remaining percentage of misfitting items was 41% (Table 28).

We shall now look at guessing as another possible source of misfit.

Guessing

In principle, the simple logistic model applies only to free response items. However, multiple choice tests could fit the model if

guessing is negligible. As we mentioned in Chapter III, Panchapakesan (1969) provided some guidelines for minimizing the effect of guessing on calibration and measurement. We shall now examine her rationale.

The most commonly used model for correcting for guessing is the random guessing model where the corrected score r' is equal to

$$r - (K-r) \Big/ (m-1)$$

where r = the number right score

K = the total number of items

m = the number of alternatives

When there are no omitted items r' is an unbiased estimate of the score an examinee would have obtained purely on the basis of knowing the answer. However, this correction is not appropriate when a subject gets an item wrong due to causes other than random guessing like misinformation or partial ignorance.

> "A somewhat more realistic model might assume random guessing after the elimination of one or more of the distractors, the number eliminated being a (probabilistic) function of the examinee's ability" (Lord and Novick, 1968, p. 309).

The guessing model used by Panchapakesan makes a provision for "intelligent" guessing. In that model, it is assumed that the number of distractors eliminated in an item is a function of the probability that a subject will get that item right. Therefore, a subject will not guess if he has a probability of one-half or greater of getting an item correct. If his probability of getting the item right lies between one-third and one-half, it is assumed that he can eliminate all but two of the distractors so his probability of getting the item right

will be a half. This stepwise process is continued till his probability becomes less than 1/m where m is the number of alternatives for the item. The model is represented as follows.

$$
\begin{aligned}
&\text{If } p \geq .5 && \text{no guessing}\\
&\text{If } .5 > p \geq .33 && \text{set } p = .5\\
&\text{If } .33 > p \geq .25 && \text{set } p = .33\\
&\quad \cdots\cdots\\
&\text{If } 1/(m-1) > p \geq 1/m && \text{set } p = 1/(m-1)\\
&\text{If } 1/m > p && \text{set } p = 1/m
\end{aligned}
$$

Simulating data on the basis of that model, Panchapakesan examined what measures could be taken to minimize the part played by guessing so that there is minimal bias in the calibration of items. She concluded that if the average ability of the calibrating sample is greater than the average difficulty of the test, it is reasonable to assume that very few subjects will resort to guessing. Table 29 shows that this is the case for the three MCAT subtests. In the simple logistic model, the average difficulty of the test is 0 by definition. Since the easinesses and the abilities are expressed on the same log scale, it can be seen that the average log ability estimate of each subtest exceeds the average log easiness. In table 29, the range of the ability estimates is also reported along with the K-R 20 estimate for each subtest.

But for practical applications, a definitive statement about the minimum ability required for a particular test so that guessing can be ignored must be made. On the basis of her guessing model and simulations, Panchapakesan (1969) suggested that for a particular test, only subjects getting a score greater than or equal to r* be used to calibrate all the

items within the test where r* is given by

$$r^* = K/m + 2\left[K\,(m-1)\,/\,m^2\right]^{\frac{1}{2}}$$

and where K = the number of items
m = the number of alternatives
K/m = the expected score solely on the basis of guessing
$K(m-1)\,/\,m^2$ = the variance of the score

The purpose of such a procedure is to focus on those subjects who have a low probability of guessing blindly among the m alternatives and thus to reduce the discrepancy in the estimation of item parameters due to guessing.

TABLE 29

AVERAGE ABILITY OF EXAMINEES FOR THE THREE MCAT SUBTESTS

	QUANTITATIVE	VERBAL	SCIENCE
Mean score	34	44	54
Log Ability*	.994	.480	.613
Minimum Log Ability	-3.381	-2.936	-2.734
Maximum Log Ability	+4.468	+4.721	+4.798
K-R 20	0.89	0.90	0.86
Average difficulty**	0.68	0.58	0.62

KEY

* log ability estimate corresponding to mean score
** average difficulty in the classical sense
K-R 20 = Kuder-Richardson Formula 20 Reliability

Table 30 gives r* for each MCAT subtest, the ability level associated with r*, and the percentage of examinees scoring below r*.

TABLE 30

NUMBER OF EXAMINEES SCORING BELOW r* IN EACH MCAT SUBTEST

	QUANTITATIVE (K = 50)	VERBAL (K = 75)	SCIENCE (K = 86)
r*	19	27	30
log ability estimate	-.600	-.658	-.721
percentage below r*	4.7%	5.7%	1.7%

Given that the average ability of the sample is greater than the average difficulty of the test and that there is a very low percentage of subjects scoring below r* in each of the three MCAT subtests, we conclude that guessing is not an important factor to explain the misfit of our data.

Furthermore, it seems as if for tests of length 50 to 100 items and $m = 4$ the log ability corresponding to an absence of guessing would be in the range -.600 to -.800. Panchapakesan (1969) noted a value of -.250 when $m = 5$ and + 1.25 when $m = 2$ for tests of length 10 to 40 items.

However, to systematically explore the presence of guessing on an item by item basis, we would want a more microscopic procedure. For items involving some guessing, Panchapakesan looked at two curves plotted against each other, the observed ICC and the theoretical ICC. It appeared that for $P < .5$ the observed points lay above the expected curve and for $P > .5$ the observed points lay below the expected curve.

Thus, for lower score groups the normal deviates were large and positive and for higher score groups the normal deviates were negative.

> "The presence of such a trend in items where guessing is taking place would suggest that the correlation between the normal deviate and the score group could be used as an index of guessing. We would expect this index to be large and negative for items where guessing is effective." (Panchapakesan, 1969, p. 108).

We actually computed such an index and identified a few patterns which we shall now discuss. The index is simply Pearson's product moment correlation coefficient. The two variables are all values of the normal deviates which exceed ± 2 and the corresponding score groups. A t statistic was used to test the significance of r, where $t = r\left[(n-2)/(1-r^2)\right]^{\frac{1}{2}}$ with n-2 degrees of freedom. The probability level chosen was .001.

Table 31 shows the items for which r is negative and significantly large in the three subtests.

TABLE 31

ITEMS WITH SIGNIFICANT NEGATIVE CORRELATION BETWEEN NORMAL DEVIATES AND SCORE GROUPS

SUBTEST	NUMBER OF ITEMS	LIST OF ITEMS
Quantitative (K = 50)	12	1,12,14,22,26,28,30,35,41,46,49,50
Verbal (K = 75)	26	10,22,24,28,29,40,42,44,47,48,50,53,54, 55,58,59,60,64,65,66,67,68,71,72,74,75
Science (K = 86)	26	4,5,7,13,14,15,16,27,28,33,35,36,37,45, 49,52,55,56,57,64,70,78,79,80,83,84

The pattern presently looked at is of the type (+,-), that is, there are more subjects than expected in the lower ability group who got the item right and less subjects than expected in the higher ability group who got the item right. This sort of pattern could reflect a guessing effect. But to be sure, one has to consider the difficulty of each item involved. Table 32 provides the distribution of items according to their respective difficulty level. It can be seen that none of the items were answered correctly by less than 20 percent of the subjects. Out of the total of 211 items, only 25 could be categorized as difficult, the rest of them being easy or of average difficulty. If we divide the items with significant negative correlation between normal deviates and score groups into three categories, easy, average, and difficult, we get the results shown in table 33. Of the 64 items listed in that table, only 12 are considered difficult ($p < .4$). Is there a guessing effect in those 12 items? It is difficult to conclude at this stage but one can reasonably assume that this set of 12 items is the one most likely to involve some guessing effect. This set represents less than 6 percent of the total pool of items. However, since an item appears easier than it really is when guessing is present, some of the items in the average difficulty category may very well contain some guessing bias. As for the rest of the 64 items represented in table 33, a (+, -) pattern when an item is easy could suggest, according to Wright (1969), indifference or careless performance. We shall look at the practical significance of such effects in the next chapter.

TABLE 32

DISTRIBUTION OF ITEMS ACCORDING TO SOME RANGES OF DIFFICULTY

RANGE OF DIFFICULTY*	QUANTITATIVE		VERBAL		SCIENCE	
	N	ITEMS	N	ITEMS	N	ITEMS
0 - 0.19	0	--	0	--	0	--
0.20 - 0.39	2	49,50	14	22,26,27,29,30, 47,49,51,52,53, 54,55,72,73	9	57,67,68,70,71, 72,73,81,84
0.40 - 0.59	15	22,29,30,32, 37,38,39,40, 41,42,43,45, 46,47,48	28	10,14,15,16,17, 18,19,20,21,23, 24,25,28,40,43, 44,45,46,48,50, 65,66,67,68,70, 71,74,75	23	13,14,15,18,46, 48,49,52,53,54, 55,58,59,60,61, 62,63,64,65,66, 69,83,86
0.60 - 0.79	17	15,17,18,19, 20,21,24,25, 26,27,28,31, 33,34,35,36, 44	19	8,9,11,12,13,34, 35,38,39,41,42, 58,59,60,61,62, 63,64,69	41	4,5,7,9,10,16, 17,22,23,25,27, 28,29,30,31,32, 33,34,35,36,37, 38,39,40,41,42, 43,44,45,47,50, 51,74,75,76,77, 78,79,80,82,85
0.80 - 1.0	16	1,2,3,4,5,6, 7,8,9,10,11, 12,13,14,16, 23	14	1,2,3,4,5,6,7, 31,32,33,36,37, 56,57	13	1,2,3,6,8,11,12, 19,20,21,24,26, 56

KEY: * difficulty index in the classical sense

TABLE 33

LEVEL OF DIFFICULTY OF ITEMS WITH SIGNIFICANT NEGATIVE CORRELATION BETWEEN NORMAL DEVIATES AND SCORE GROUPS

SUBTEST	DIFFICULTY		LIST OF ITEMS
	LEVEL	NUMBER	
Quantitative (N = 12)	Easy	6	1,12,14,26,28,35
	Average	4	22,30,41,46
	Difficult	2	49,50
Verbal (N = 26)	Easy	5	42,58,59,60,64
	Average	14	10,24,28,40,44,48,50,65,66,67, 68,71,74,75
	Difficult	7	22,29,47,53,54,55,72
Science (N = 26)	Easy	15	4,5,7,16,27,28,33,35,36,37, 45,56,78,79,80
	Average	8	13,14,15,49,52,55,64,83
	Difficult	3	57,70,84

Speed

The simple logistic model applies only to power tests. Sedlacek (1967) provides some data on the 1966 group of MCAT examinees. For that year, the percentage of examinees who did not finish the MCAT in the prescribed time was 16.21 for the Quantitative subtest, 3.18 for the Verbal subtest, and 0.12 for the Science subtest. We ran a classical item analysis on a random sample of 2000 subjects drawn from our overall 1972 sample and found the same trend in the bottom 27 percent of the subjects. In this group, the percentage of omitted responses showed that there was a speed factor involved in the three subtests which was more important for the Quantitative subtest than for the other two subtests. How important this speed factor is on item calibration and

on measurement, we shall see in Chapter VI. To identify which specific items may be affected by a speed factor is not an easy task. Such a factor could be confounded with guessing for instance. Therefore, a (+,-) type of pattern could be interpreted differently. It could mean that even though some of the subjects of lower ability might not have had a chance to attempt an item, some of them might have had sufficient time to guess blindly at that item. The resulting pattern would then look as if guessing was the main source of misfit. But this would be plausible only for the last items in the test. A more likely pattern for reflecting a speed factor would be of the (-,+) type. This pattern means that there are less people than expected in the lower ability group and more subjects than expected in the higher ability group. It is indeed reasonable to expect that those who cannot attempt an item and thus cannot guess at it would be underrepresented in the data matrix. This is more likely to be the case for subjects of low ability. On the other hand, if pressed by time limits, subjects of high ability may resort to guessing and the resulting pattern would be (-,+). But again such an hypothesis would be viable only for the last items in the test.

Using this rationale on a very tentative basis, we looked at the items for which the correlation between the normal deviates and the score groups was significantly positive, thus reflecting a (-,+) pattern. We divided the items into three categories according to their rank in the test. The data are presented in table 34. There are 54 items with such a significant positive correlation. If we can assume that speed would not show up in the first two-thirds of the items, then only 13 items are likely to be affected by a speed factor alone.

As for the 41 remaining items, their source of misfit must be attributed to factors other than speed and guessing.

Looking back at table 33, one may conclude on the basis of the same rationale that some items are very likely to be affected by both factors, that is, guessing and speed. Those items are items 49 and 50 in the Quantitative subtest and items 53,54,55, and 72 in the Verbal subtest.

The procedure followed so far does not permit a positive identification of causes of misfit. However, it does allow us to eliminate the most improbable sources of misfit for a given set of items.

Before concluding this chapter, we shall now summarize the results of our investigations.

TABLE 34

ITEMS WITH SIGNIFICANT POSITIVE CORRELATION BETWEEN NORMAL DEVIATES AND SCORE GROUPS

SUBTEST	ITEM RANK		LIST OF ITEMS
	LEVEL	NUMBER	
Quantitative (N = 10)	1-16	4	6,10,11,13
	17-32	2	19,25
	33-50	4	33,34,37,40
Verbal (N = 23)	1-25	11	3,4,5,6,7,9,12,13,14,15,25
	26-50	7	31,32,34,35,36,41,49
	51-75	5	52,56,57,61,63
Science (N = 21)	1-28	12	2,9,10,11,12,17,19,21,23,24, 25,26
	29-56	5	31,34,38,40,50
	57-86	4	59,62,75,82

Sources of misfit in the MCAT test

In this chapter, we explored the sources of misfit in the three MCAT subtests. To do so, three statistics were used: the item chi-square, the standardized difference of the slope from unity, and the correlation between normal deviates and score groups. Our purpose was to locate the lack of fit in the data, that is, to identify the items most likely responsible for the overall misfit of the logistic model to the MCAT test. We then investigated some of the usual causes of misfit in real data: item discrimination, guessing, and speed. After eliminating mis-scoring as a possible explanation, we conclude that those three factors do play a role in the misfit of our data, the practical relevance of which shall be analyzed in the next chapter. However, before proceeding any further, we need to summarize the results obtained so far. Tables 35,36,37 provide the data for each subtest. The items are presented according to their rank in the test for easy reference. The second and third columns show the mean square value and its probability. An asterisk indicates the items that fit the model according to the item chi-square statistic. The fourth and fifth columns give the magnitude of the slope and its significance. An asterisk indicates the items for which the standardized difference of the slope from unity, adjusted for small standard errors, is not significant. The sixth and seventh columns provide the value of the correlation coefficient between normal deviates and score groups and its statistical significance. As asterisk shows the items for which r is not significant. One should note here that the sign of the coefficient is more important than its magnitude. Sometimes the degrees of freedom are so small that even values exceeding .90 are not significant. Finally, the eighth column indicates the items

which show misfit on all three criteria. Those items should definitely be the worst items in the tests. There are 12 such items in the Quantitative subtest, 29 in the Verbal subtest, and 25 in the Science subtest. This set of items totals 66 and represents 31 percent of the total pool.

If we look at r, we notice that t is not significant for all but two items for which the item chi-square is not significant (item 17 in Verbal and item 45 in Science). In these two cases however, the significance of r is spurious because of the small number of score groups involving normal deviates exceeding $\pm$ 2(9 and 6 respectively). We notice also that r does well in depicting items for which the value of the slope is less than 0.6. It is not significant for only 13 such items. However, for 7 of those items, the item chi-square is not significant either, and for the remaining 6 items, the mean square value does not exceed 2.31. From the evidence gathered so far, we conclude that the three criteria should be used together since no single one of them seems to get at all sources of misfit at once and each one of them appears to point at different causes of misfit.

Hence, to interpret tables 35,36, and 37, we suggest the following procedure. If the item chi-square is not significant, don't look any further, the item fits. This conclusion stems from the fact that this criterion is the most sensitive of all three. To identify the best set of items, sort out the items that show fit on all three criteria. There are 20 such items in the three subtests combined. To identify the worst set of items, pick out those which show misfit on all three criteria. There are 66 of those in the MCAT subtests. For this set of items, we offer the following tentative interpretation. There are

probably many reasons responsible for the lack of fit or there is a single reason which is so important that the item looks terrible on all counts, bad wording for instance. For the rest of the items, one must proceed by elimination. For example, an item does not fit according to the item chi-square statistic but does fit with the other two criteria (Q.A., item 2). In such a case, one can only conclude that either the item fits (the item chi-square is too sensitive) or the cause of misfit is not related to item discrimination, nor to guessing, nor to speed. Another example would be an item that shows fit only with r (Q.A., item 4). Here, guessing and speed can be excluded. Similarly, if an item shows fit only with S, item discrimination can be eliminated in explaining the misfit (Q.A., item 6). If an item fits according to both chi-square and S or only with S, one must then look at the sign of r. If r is positive and the item ranks in the last third of the test, a speed factor is a likely hypothesis. If r is negative and the item is difficult then guessing is a possible cause of misfit.

In Chapter VI, the overall sample will be divided into different subgroups, as explained in Chapter IV, and scoring tables computed from each sub-group will be compared. We shall then examine how the procedures of item fit described here relate to indicators of fit at the test level, and thus, what can be learned from such a relation that would be useful for further studies of model-data fit.

TABLE 35

QUANTITATIVE ABILITY SUBTEST - SUMMARY RESULTS

ITEM	MEAN SQUARE	p ≥.001	SLOPE	|S|* < 3	r	p > .001	MISFIT
1	6.97		0.561		-.90		*
2	3.79		0.789	*	-.68	*	
3	5.86		0.709	*	-.65	*	
4	2.98		0.644		-.57	*	
5	1.81		0.925	*	.50	*	
6	5.84		1.038	*	.91		
7	1.90		0.954	*	.89	*	
8	1.95		0.799	*	.12	*	
9	1.40	*	0.872	*	.96	*	
10	7.96		1.168	*	.88		
11	3.92		0.978	*	.93		
12	4.83		0.561		-.88		*
13	5.92		1.046	*	.87		
14	6.91		0.488		-.90		*
15	1.75	*	0.622		-.92	*	
16	1.26	*	0.696		-.53	*	
17	2.12		0.840	*	.50	*	
18	0.85	*	0.820	*	.00	*	
19	5.13		0.940	*	.90		
20	1.72	*	0.650		-.92	*	
21	1.29	*	0.714	*	-.33	*	
22	15.24		0.329		-.88		*
23	2.74		0.824	*	.53	*	
24	1.37	*	0.754	*	.00	*	
25	12.38		1.031	*	.88		
26	7.91		0.459		-.88		*
27	1.29	*	0.774	*	.63	*	
28	3.05		0.512		-.92		*
29	3.64		0.754	*	-.27	*	
30	6.46		0.469		-.91		*
31	1.46	*	0.704	*	-.68	*	
32	1.95		0.787	*	.33	*	
33	3.79		0.786	*	.88		
34	9.06		1.176	*	.88		
35	2.51		0.532		-.96		*
36	3.79		0.717	*	-.30	*	
37	3.90		0.769	*	.84		
38	2.09		0.660		-.15	*	
39	3.30		0.626		-.27	*	
40	3.05		0.787	*	.84		
41	6.65		0.531		-.77		*
42	1.00	*	0.747	*	.00	*	
43	12.16		0.809	*	.33	*	

TABLE 35 - Continued

ITEM	MEAN SQUARE	$p \geq .001$	SLOPE	$\lvert S \rvert^* < 3$	r	$p > .001$	MISFIT
44	5.75		0.891	*	.62	*	
45	4.69		0.764	*	.29	*	
46	13.51		0.461		-.88		*
47	4.34		0.650		-.37	*	
48	6.99		0.749	*	.10	*	
49	7.97		0.603		-.62		
50	31.88		0.335		-.83		

TABLE 36

VERBAL ABILITY SUBTEST - SUMMARY RESULTS

ITEM	MEAN SQUARE	$p \geq .001$	SLOPE	$\lvert S \rvert^* < 3$	r	$p > .001$	MISFIT
1	2.72		1.177	*	.17	*	
2	2.00		0.915	*	.53	*	
3	5.30		1.169	*	.88		*
4	16.02		1.538		.74		
5	8.53		1.405	*	.85		
6	7.48		1.185	*	.82		
7	4.04		1.008	*	.70		
8	1.11	*	0.827	*	.32	*	
9	4.66		0.952	*	.74		*
10	2.62		0.443		-.98		
11	1.41	*	0.684		-.57	*	
12	10.87		1.235	*	.83		
13	5.93		1.021	*	.88		
14	4.29		0.944	*	.58		
15	4.30		0.742	*	.69		
16	4.03		0.660		-.08	*	
17	1.57	*	0.653		.89		
18	5.17		0.830	*	.50	*	
19	2.27		0.712	*	.01	*	
20	4.54		0.621		-.26	*	
21	3.85		0.684		-.11	*	
22	3.99		0.451		-.75		*
23	3.30		0.801	*	.65	*	
24	2.98		0.489		-.95		*
25	4.33		0.902	*	.65		
26	5.61		0.708	*	-.15	*	
27	1.58	*	0.624		-.23	*	
28	3.37		0.557		-.90		*
29	5.79		0.537		-.64		*
30	6.23		0.633		.05	*	
31	4.75		0.981	*	.61		
32	6.70		1.370	*	.89		
33	2.24		0.912	*	.19	*	
34	7.30		1.023	*	.87		
35	21.90		1.524		.76		*
36	16.42		1.448		.80		*
37	5.61		0.698		.19	*	
38	2.13		0.783	*	.61	*	
39	3.52		0.713	*	-.15	*	
40	3.53		0.554		-.91		*
41	8.33		0.948	*	.76		
42	6.24		0.429		-.87		*
43	0.94	*	0.763	*	.20	*	

TABLE 36 - Continued

ITEM	MEAN SQUARE	p ≥.001	SLOPE	|S|* < 3	r	p > .001	MISFIT
44	6.33		0.409		-.83		*
45	2.28		0.487		-.68	*	
46	1.74		0.549		-.38	*	
47	9.53		0.545		-.73		*
48	5.68		0.364		-.85		*
49	9.38		0.930	*	.52		
50	4.06		0.580		-.77		*
51	4.47		0.705	*	-.34	*	
52	3.64		0.903	*	.62		
53	4.54		0.389		-.93		*
54	9.90		0.491		-.86		*
55	6.26		0.378		-.93		*
56	4.80		1.280	*	.86		
57	5.15		1.065	*	.87		
58	29.08		0.133		-.91		*
59	14.04		0.356		-.91		*
60	4.81		0.549		-.91		*
61	2.75		0.989	*	.78		
62	1.72		0.610		-.67	*	
63	5.21		1.002	*	.87		
64	7.84		0.526		-.90		*
65	10.79		0.338		-.88		*
66	4.71		0.419		-.90		*
67	18.02		0.177		-.90		*
68	4.52		0.517		-.92		*
69	2.03		0.908	*	.83	*	
70	3.08		0.624		-.41	*	
71	7.91		0.304		-.89		*
72	7.49		0.366		-.91		*
73	1.57	*	0.591		.20	*	
74	2.33		0.557		-.82		*
75	20.40		0.163		-.89		*

TABLE 37

SCIENCE SUBTEST - SUMMARY RESULTS

ITEM	MEAN SQUARE	$p \geq .001$	SLOPE	$\lvert S \rvert$ * < 3	r	p > .001	MISFIT
1	1.58	*	0.849	*	-.98	*	
2	2.52		1.220	*	.95		
3	0.82	*	0.950	*	.00	*	
4	1.80		0.686		-.87		*
5	2.93		0.542		-.87		*
6	1.43	*	1.001	*	.10	*	
7	4.64		0.446		-.90		*
8	1.83		0.672		-.43	*	
9	5.58		1.090	*	.65		
10	4.73		1.037	*	.74		
11	2.27		1.100	*	.92		
12	3.70		1.109	*	.89		
13	3.23		0.416		-.91		*
14	6.66		0.327		-.89		*
15	12.67		0.313		-.84		*
16	3.46		0.498		-.88		*
17	6.41		0.918	*	.58		
18	3.40		0.622		.43	*	
19	1.92		0.968	*	.90		
20	1.13	*	0.930	*	-.93	*	
21	6.06		1.317	*	.88		
22	1.50	*	0.861	*	-.19	*	
23	5.67		0.905	*	.60		
24	3.15		0.829	*	.69		
25	2.49		1.035	*	.94		
26	4.15		1.029	*	.91		
27	3.52		0.438		-.89		*
28	1.62		0.535		-.90		*
29	1.47	*	0.571		-.92	*	
30	1.65		0.797	*	.06	*	
31	2.36		0.906	*	.94		
32	1.40	*	0.667		-.81	*	
33	2.96		0.557		-.89		*
34	4.79		1.144	*	.86		
35	4.74		0.341		-.85		*
36	2.13		0.437		-.84		*
37	2.15		0.556		-.88		*
38	2.98		0.945	*	.71		
39	1.25	*	0.649		.18	*	
40	2.51		1.002	*	.75		
41	1.03	*	0.754	*	-.08	*	
42	2.00		0.426		-.45	*	
43	2.31		0.397		-.75	*	

TABLE 37-Continued

ITEM	MEAN SQUARE	p ≥ .001	SLOPE	|S|* < 3	r	p > .001	MISFIT
44	3.00		0.774	*	.49	*	
45	1.30	*	0.575		-.98		
46	1.56	*	0.453		-.51	*	
47	1.17	*	0.544		-.50	*	
48	0.98	*	0.656		-.43	*	
49	3.54		0.369		-.84		*
50	2.86		0.939	*	.92		
51	1.19	*	0.575		-.87	*	
52	1.61		0.521		-.98		*
53	1.00	*	0.642		.25	*	
54	1.97		0.730	*	.46	*	
55	10.49		0.114		-.84		*
56	4.70		0.487		-.89		*
57	7.07		0.374		-.89		*
58	1.46	*	0.456		-.83	*	
59	9.54		0.916	*	.83		
60	1.95		0.447		-.35	*	
61	1.57	*	0.738	*	.88	*	
62	4.59		1.008	*	.85		
63	1.50	*	0.515		-.39	*	
64	3.87		0.385		-.90		*
65	2.11		0.710	*	.14	*	
66	1.05	*	0.678		.00	*	
67	3.92		0.746	*	.37	*	
68	2.12		0.574		-.57	*	
69	0.95	*	0.721	*	.00	*	
70	3.56		0.450		-.92		*
71	2.12		0.734	*	-.40	*	
72	1.59	*	0.737	*	-.03	*	
73	2.85		0.642		-.03	*	
74	0.93	*	0.948	*	.00	*	
75	2.28		0.844	*	.92		
76	0.85	*	0.730	*	.00	*	
77	1.54	*	0.659		-.39	*	
78	3.07		0.486		-.89		*
79	3.29		0.552		-.70		*
80	2.69		0.359		-.91		*
81	1.49	*	0.613		.77	*	
82	2.73		0.816	*	.85		
83	2.17		0.468		-.90		*
84	7.68		0.258		-.90		*
85	1.09	*	0.763	*	-.24	*	
86	1.10	*	0.531		-.87	*	

CHAPTER VI

EFFECTS OF MISFIT ON CALIBRATION AND MEASUREMENT

The main advantage of any simulation procedure over the use of real data is that of providing a rigorous way of assessing the effect of departures from the model on item calibration and person measurement. One can simply examine the difference between the generating parameters and the estimated parameters to verify the robustness of the model to a violation of its assumptions. With real data however one never knows for sure how bad an item really is. The "true" state of affairs is unknown. In such a case one reasonable method of evaluating fit is to compare the calibration results over extremely different samples of people. Using that procedure, we shall now look at the relative divergence of scoring tables based on ability estimates and the relative instability of item parameters for the three MCAT subtests. It should be noted that none of the bad items identified in Chapter V were thrown out for this analysis so that the effects of misfit could be fully appreciated.

To proceed further in exploring the fit of our data, we divided the total sample into eight subgroups and ran eight analyses per subtest, one for each cell. The split creating the 8 groups was based on two variables, total test score (above and below the median) and parents' income level (four categories). The corresponding subgroups are given in table 38 along with their respective sample sizes. It should be noted that even after such a split the sample sizes remain quite large, the smallest being 325 (Science, AMI). Thus, the relative instability of the log estimates could hardly be attributed to the size of the samples.

The choice of the two variables is arbitrary. The purpose is

TABLE 38

FREQUENCY OF SUBJECTS WHO SCORED ABOVE AND BELOW THE MEDIAN IN EACH OF FOUR DIFFERENT INCOME LEVEL GROUPS FOR THE THREE MCAT SUBTESTS

SUBTEST	INCOME LEVEL	ABOVE MEDIAN	BELOW MEDIAN	TOTAL
QUANTITATIVE	I	337	693	1030
	II	1045	1575	2620
	III	3111	3206	6317
	IV	2613	2187	4800
	TOTAL	7106	7661	14767
VERBAL	I	374	656	1030
	II	1200	1420	2620
	III	3221	3096	6317
	IV	2854	1946	4800
	TOTAL	7649	7118	14767
SCIENCE	I	325	705	1030
	II	1107	1513	2620
	III	3203	3114	6317
	IV	2648	2152	4800
	TOTAL	7283	7484	14767

Key:

INCOME LEVEL I : < $5,000
II : $5,000 - $9,999
III: $10,000 - $19,999
IV : ≥ $20,000

QUANTITATIVE ABOVE MEDIAN: scores > 35
BELOW MEDIAN: scores ≤ 35

VERBAL ABOVE MEDIAN: scores > 44
BELOW MEDIAN: scores ≤ 44

SCIENCE ABOVE MEDIAN: scores > 54
BELOW MEDIAN scores ≤ 54

only to make the contrast as wide as possible. Wright (1968) used the smart-dumb split in his investigation of the Law School Admission Test to demonstrate the independence of test calibration on the abilities of 976 beginning law students. In this study, to increase even more the contrast, we added a very often used SES variable, that of parents' income.

Relative divergence of scoring tables

To each score group corresponds a log ability estimate. If the model represents the data well, for each score group the log ability estimate obtained from one sample should be identical to the log ability estimate obtained from any other sample. This is the basic requirement of a person-free test calibration method. A first series of scoring tables based on the split presented in table 38 was computed. The data are shown in Appendices A1, A2, and A3. The first column contains the score groups. The second column shows the estimates obtained from the total sample. The following eight columns present the log ability estimates derived from the sub-samples. Within each score group, the estimates are strikingly close to one another the difference varying between .01 and .20 for non empty score groups. Are they close enough to consider them identical? For practical purposes they surely are but it would help to rest such a decision on a statistical test which would take into account the standard errors of the log ability estimates. These standard errors are presented in Appendices A4, A5, and A6. They are in general quite small which suggests the precision of the estimates. Furthermore, they are so close to one another that we can suppose that there is no gain in precision to be

expected with samples of size greater than 325 (Science, AMI). Table 38 indicates that the range of sample sizes used in this series of investigations goes from 325 to 14,767. The difference between the magnitudes of the standard errors across the sub-groups does not exceed .01.

In order to test the degree of divergence between the eight log ability estimates within each score group, the following statistic was computed: $\chi^2 = \Sigma w_j(\hat{\Theta}_j - \hat{\Theta}_o)^2$.
This weighted average deviation is distributed approximately as a chi-square with j-1 degrees of freedom, where j varies over the sub-groups $j = 1, 2, \ldots 8$,

$\hat{\Theta}_j$ is the estimated log ability for sub-group j

and $\hat{\Theta}_o = \sum_{j=1}^{j} w_j\hat{\Theta}_j \Big/ \sum_{j=1}^{j} w_j$

where $w_j = 1/S.E._{(\hat{\Theta}_j)}^2$

and $S.E._{(\hat{\Theta}_j)}$ is the estimated standard error of the log ability estimate.

If the hypothesis of no difference is true, the sum of the weighted squared difference between the within-group log ability estimate $\hat{\Theta}_j$ and a weighted average of all log ability estimates $\hat{\Theta}_o$ should be zero within each score group r.

With such a test, denoted as χ_r^2, it becomes possible to assess the effects of all sources of item misfit on the ability estimates for each score group. Tables 39,40, and 41 provide the value of χ_r^2 for the score groups of the three MCAT subtests. Most of these values are smaller than 1.0. To be significant at the .001 level they must exceed 24.322

TABLE 39

QUANTITATIVE ABILITY - DEGREE OF DIVERGENCE BETWEEN SCORING TABLES COMPUTED FROM SUB-GROUPS OF TABLE 38

SG	$\hat{\theta}_o$	χ^2_r	SG	$\hat{\theta}_o$	χ^2_r
1	-4.6429	.0259	26	.1467	.0101
2	-3.8841	.0416	27	.2515	.0106
3	-3.4144	.0515	28	.3564	.0111
4	-3.0633	.0568	29	.4616	.0110
5	-2.7780	.0611	30	.5677	.0117
6	-2.5342	.0610	31	.6751	.0129
7	-2.3191	.0612	32	.7838	.0153
8	-2.1250	.0609	33	.8947	.0174
9	-1.9472	.0589	34	1.0074	.0193
10	-1.7824	.0536	35	1.1233	.0227
11	-1.6252	.0535	36	1.2427	.0263
12	-1.4812	.0492	37	1.3665	.0305
13	-1.3423	.0446	38	1.4954	.0347
14	-1.2090	.0389	39	1.6308	.0388
15	-1.0812	.0369	40	1.7737	.0449
16	- .9572	.0329	41	1.9262	.0496
17	- .8371	.0299	42	2.0909	.0553
18	- .7207	.0257	43	2.2710	.0627
19	- .6062	.0226	44	2.4718	.0683
20	- .4945	.0198	45	2.7005	.0731
21	- .3828	.0175	46	2.9707	.0757
22	- .2765	.0151	47	3.3058	.0752
23	- .1693	.0130	48	3.7595	.0669
24	- .0635	.0119	49	4.5027	.0453
25	.0418	.0105			

Key: SG: score group
$\hat{\theta}_o$: weighted average of log ability estimates
χ^2_r: within-score group chi-square with 7 degrees of freedom

TABLE 40

VERBAL ABILITY - DEGREE OF DIVERGENCE BETWEEN SCORING TABLES COMPUTED FROM SUB-GROUPS OF TABLE 38

SG	$\hat{\theta}_o$	χ^2_r	SG	$\hat{\theta}_o$	χ^2_r
1	-5.1473	.5983	38	.1091	.1738
2	-4.3561	.8315	39	.1759	.2083
3	-3.8736	.6761	40	.2421	.2426
4	-3.5204	1.0373	41	.3088	.2814
5	-3.2392	1.0821	42	.3756	.3150
6	-3.0034	1.0983	43	.4419	.3539
7	-2.7990	1.0882	44	.5088	.3916
8	-2.6186	1.0611	45	.5764	.4284
9	-2.4552	1.0193	46	.6441	.4598
10	-2.3060	.9643	47	.7122	.4989
11	-2.1624	.9018	48	.7813	.5331
12	-2 0390	.8301	49	.8512	.5650
13	-1.9187	.7500	50	.9221	.5987
14	-1.8044	.6768	51	.9938	.6282
15	-1.6959	.5975	52	1.0667	.6514
16	-1.5922	.5260	53	1.1413	.6753
17	-1.4930	.4483	54	1.2172	.7014
18	-1.3975	.3831	55	1.2951	.7196
19	-1.3056	.3184	56	1.3753	.7336
20	-1.2166	.2600	57	1.4576	.7520
21	-1.1304	.2058	58	1.5426	.7668
22	-1.0464	.1576	59	1.6308	.7646
23	- .9645	.1167	60	1.7226	.7677
24	- .8847	.0834	61	1.8184	.7654
25	- .8068	.0559	62	1.9194	.7612
26	- .7303	.0354	63	2.0260	.7569
27	- .6554	.0198	64	2.1393	.7392
28	- .5815	.0112	65	2.2605	.7246
29	- .5090	.0079	66	2.3920	.7024
30	- .4374	.0101	67	2.5357	.6801
31	- .3669	.0179	68	2.6957	.6517
32	- .2974	.0292	69	2.8761	.6240
33	- .2282	.0459	70	3.0857	.5994
34	- .1598	.0656	71	3.3371	.5805
35	- .0919	.0886	72	3.6552	.5856
36	- .0246	.1162	73	4.0944	.6531
37	.0424	.1642	74	4.7610	.0962

TABLE 41

SCIENCE - DEGREE OF DIVERGENCE BETWEEN
SCORING TABLES COMPUTED FROM SUB-GROUPS OF TABLE 38

SG	$\hat{\theta}_o$	χ^2_r	SG	$\hat{\theta}_o$	χ^2_r
1	-4.8609	.0240	44	.0585	.0093
2	-4.1404	.0445	45	.1132	.0102
3	-3.7085	.0621	46	.1682	.0117
4	-3.3945	.0766	47	.2233	.0134
5	-3.1455	.0886	48	.2786	.0162
6	-2.9377	.0977	49	.3340	.0175
7	-2.7586	.1052	50	.3900	.0202
8	-2.5999	.1114	51	.4463	.0233
9	-2.4574	.1144	52	.5030	.0253
10	-2.3275	.1160	53	.5601	.0286
11	-2.2076	.1184	54	.6178	.0323
12	-2.0963	.1166	55	.6766	.0358
13	-1.9918	.1143	56	.7356	.0392
14	-1.8933	.1126	57	.7957	.0427
15	-1.8001	.1098	58	.8566	.0463
16	-1.7110	.1060	59	.9187	.0488
17	-1.6260	.1009	60	.9815	.0521
18	-1.5444	.0955	61	1.0461	.0562
19	-1.4658	.0906	62	1.1117	.0593
20	-1.3898	.0843	63	1.1788	.0630
21	-1.3162	.0786	64	1.2479	.0658
22	-1.2450	.0722	65	1.3186	.0691
23	-1.1754	.0679	66	1.3917	.0723
24	-1.1077	.0616	67	1.4673	.0740
25	-1.0417	.0565	68	1.5450	.0767
26	- .9769	.0510	69	1.6260	.0812
27	- .9132	.0454	70	1.7104	.0829
28	- .8510	.0406	71	1.7982	.0843
29	- .7900	.0354	72	1.8907	.0876
30	- .7295	.0305	73	1.9881	.0878
31	- .6701	.0279	74	2.0916	.0883
32	- .6114	.0234	75	2.2020	.0905
33	- .5534	.0203	76	2.3205	.0919
34	- .4959	.0171	77	2.4493	.0940
35	- .4390	.0147	78	2.5909	.0939
36	- .3827	.0126	79	2.7475	.0946
37	- .3270	.0107	80	2.9258	.0951
38	- .2711	.0093	81	3.1321	.0959
39	- .2157	.0087	82	3.3800	.0960
40	- .1608	.0081	83	3.6928	.0944
41	- .1057	.0080	84	4.1250	.0870
42	- .0508	.0081	85	4.8469	.0652
43	.0038	.0082			

with 7 degrees of freedom. Hence, the log ability estimates for each score group are statistically identical across the eight sub-groups. This is the case even for score groups for which there are no subjects. That sounds impossible but it follows directly from the item analysis model used for these investigations. Even with the same total score, persons differ in those items on which they succeed. When the calibration sample is large, these differences can be used to calibrate the items, and hence, the test over its entire range of possible scores, even though only one score has actually been observed (Wright, 1968).

Thus, we conclude that the degree of item misfit noted in the preceding chapter has no effect whatsoever on ability measurement. The three MCAT subtests can be used with confidence. Each subtest could have been calibrated on any one of the eight sub-groups described earlier and applied to the other sub-groups without fear of being unfair to any given group of subjects. The estimated ability would be the same for each score group, and therefore, any decision made on the basis of these estimated log abilities would be the same across sub-groups. This seems a proper definition of test fairness! It is at least a good demonstration of what is meant by a sample-free measurement model.

These same investigations were additionally carried out on a different set of sub-samples. The overall sample was divided according to the subjects' racial background. There was a total of 12,599 usable records for that split. Table 42 shows the breakdown.

A second series of scoring tables based on this racial split is presented in Appendices A7, A8, and A9. The standard errors of this

TABLE 42

NUMBER OF SUBJECTS IN THREE DIFFERENT RACIAL GROUPS TAKING THE THREE MCAT SUBTESTS

RACIAL BACKGROUND	FREQUENCY
White	10,685
Black	626
Other	1,288
Total	12,599

new set of log ability estimates are given in Appendices A10, A11, and A12.

Again, we can see that the log ability estimates are very close to one another, the difference varying from .005 to .008 for the Quantitative, from .002 to .048 for the Verbal, and from .003 to .005 for the Science subtest. Tables 43,44, and 45 give the value of χ^2_r for the score groups of the three subtests. All of these values are smaller than 1.0. With 2 degrees of freedom and a probability level of .001, the critical χ^2 value is 13.816. Thus, it appears clearly that the calibration of the three MCAT subtests is not affected by the racial background of the medical school applicants. Such a conclusion reinforces the notion that the measurement process is made up of two stages and that, under the simple logistic model, these two stages are independent of each other.

For practical applications, we must however reconcile the kind of fit demonstrated at the measurement phase with the fit sought at the item calibration stage. It is satisfying to note that the many possible sources of item misfit discussed in the preceding chapter do not seem to

TABLE 43

QUANTITATIVE ABILITY - DEGREE OF DIVERGENCE BETWEEN SCORING TABLES COMPUTED FROM SUB-GROUPS OF TABLE 42

SG	$\hat{\theta}_o$	χ^2_r	SG	$\hat{\theta}_o$	χ^2_r
1	-4.5556	.0093	26	.1379	.0009
2	-3.8021	.0150	27	.2402	.0013
3	-3.3371	.0183	28	.3425	.0019
4	-2.9911	.0200	29	.4452	.0025
5	-2.7104	.0201	30	.5491	.0034
6	-2.4711	.0197	31	.6541	.0044
7	-2.2608	.0185	32	.7604	.0055
8	-2.0711	.0172	33	.8688	.0063
9	-1.8978	.0155	34	.9794	.0072
10	-1.7372	.0136	35	1.0920	.0084
11	-1.5865	.0119	36	1.2104	.0093
12	-1.4446	.0101	37	1.3320	.0109
13	-1.3093	.0084	38	1.4587	.0117
14	-1.1800	.0070	39	1.5913	.0131
15	-1.0557	.0057	40	1.7320	.0137
16	- .9357	.0042	41	1.8820	.0148
17	- .8191	.0033	42	2.0440	.0153
18	- .7055	.0023	43	2.2216	.0158
19	- 5948	.0017	44	2.4193	.0159
20	- .4862	.0009	45	2.6446	.0158
21	- .3792	.0005	46	2.9113	.0152
22	- .2739	.0003	47	3.2426	.0134
23	- .1699	.0002	48	3.6913	.0106
24	- .0670	.0003	49	4.4281	.0064
25	.0359	.0005			

Key: SG: score group

$\hat{\theta}_o$: weighted average of log ability estimates

χ^2_r: within-score group chi-square with 2 degrees of freedom

TABLE 44

VERBAL ABILITY - DEGREE OF DIVERGENCE BETWEEN
SCORING TABLES COMPUTED FROM SUB-GROUPS OF TABLE 42

SG	$\hat{\theta}_o$	χ^2_r	SG	$\hat{\theta}_o$	χ^2_r
1	-4.8876	.1082	38	.0710	.0204
2	-4.1263	.1448	39	.1346	.0252
3	-3.6652	.1567	40	.1976	.0300
4	-3.3288	.1605	41	.2608	.0344
5	-3.0623	.1578	42	.3242	.0392
6	-2.8392	.1540	43	.3882	.0431
7	-2.6470	.1480	44	.4525	.0482
8	-2.4772	.1409	45	.5167	.0535
9	-2.3245	.1316	46	.5817	.0576
10	-2.1843	.1224	47	.6474	.0623
11	-2.0555	.1129	48	.7138	.0682
12	-1.9359	.1018	49	.7807	.0720
13	-1.8231	.0923	50	.8494	.0767
14	-1.7172	.0818	51	.9184	.0806
15	-1.6161	.0716	52	.9894	.0840
16	-1.5197	.0622	53	1.0610	.0870
17	-1.4278	.0531	54	1.1351	.0889
18	-1.3394	.0455	55	1.2109	.0930
19	-1.2536	.0369	56	1.2885	.0939
20	-1.1711	.0299	57	1.3689	.0955
21	-1.0909	.0242	58	1.4519	.0974
22	-1.0130	.0187	59	1.5383	.0978
23	- .9369	.0140	60	1.6277	.0980
24	- .8627	.0098	61	1.7213	.0963
25	- .7895	.0064	62	1.8197	.0954
26	- .7186	.0040	63	1.9238	.0947
27	- .6484	.0020	64	2.0348	.0908
28	- .5795	.0010	65	2.1535	.0869
29	- .5116	.0006	66	2.2821	.0825
30	- .4447	.0009	67	2.4232	.0767
31	- .3781	.0017	68	2.5795	.0710
32	- .3125	.0030	69	2.7562	.0641
33	- .2479	.0048	70	2.9607	.0567
34	- .1833	.0072	71	3.2057	.0472
35	- .1194	.0108	72	3.5158	.0370
36	- .0558	.0130	73	3.9428	.0257
37	.0077	.0170	74	4.6575	.0135

TABLE 45

SCIENCE - DEGREE OF DIVERGENCE BETWEEN
SCORING TABLES COMPUTED FROM SUB-GROUPS OF TABLE 42

SG	$\hat{\theta}_o$	χ^2_r	SG	$\hat{\theta}_o$	χ^2_r
1	-4.7962	.0050	44	.0590	.0002
2	-4.0775	.0090	45	.1126	.0001
3	-3.6474	.0120	46	.1659	.0003
4	-3.3351	.0145	47	.2199	.0004
5	-3.0884	.0165	48	.2736	.0007
6	-2.8824	.0180	49	.3279	.0010
7	-2.7047	.0191	50	.3829	.0014
8	-2.5480	.0204	51	.4375	.0021
9	-2.4071	.0208	52	.4928	.0024
10	-2.2791	.0221	53	.5485	.0032
11	-2.1607	.0220	54	.6052	.0043
12	-2.0511	.0223	55	.6621	.0050
13	-1.9484	.0225	56	.7201	.0058
14	-1.8511	.0217	57	.7788	.0071
15	-1.7594	.0216	58	.8381	.0080
16	-1.6724	.0212	59	.8987	.0095
17	-1.5883	.0208	60	.9607	.0104
18	-1.5088	.0206	61	1.0234	.0120
19	-1.4318	.0195	62	1.0873	.0129
20	-1.3567	.0187	63	1.1530	.0146
21	-1.2848	.0179	64	1.2210	.0154
22	-1.2151	.0172	65	1.2900	.0171
23	-1.1470	.0163	66	1.3614	.0187
24	-1.0804	.0145	67	1.4349	.0202
25	-1.0161	.0143	68	1.5113	.0210
26	- .9524	.0126	69	1.5907	.0224
27	- .8904	.0116	70	1.6729	.0229
28	- .8294	.0106	71	1.7597	.0241
29	- .7694	.0096	72	1.8503	.0250
30	- .7104	.0087	73	1.9459	.0258
31	- .6525	.0078	74	2.0466	.0264
32	- .5954	.0069	75	2.1553	.0267
33	- .5385	.0060	76	2.2720	.0267
34	- .4825	.0052	77	2.3980	.0269
35	- .4268	.0040	78	2.5370	.0262
36	- .3718	.0034	79	2.6910	.0244
37	- .3172	.0030	80	2.8664	.0233
38	- .2628	.0022	81	3.0691	.0211
39	- .2089	.0017	82	3.3128	.0183
40	- .1549	.0013	83	3.6214	.0148
41	- .1016	.0008	84	4.0470	.0107
42	- .0476	.0006	85	4.7608	.0059
43	.0053	.0004			

affect the fairness of the decisions reached about the examinees.

But our goal in measurement is to ensure consistency over time. Would a different set of items administered to a different group of subjects yield comparable results? To achieve such a goal, the log easiness estimates computed from different samples should also be close to one another for any given test. Obviously, such a within-test stability is a prerequisite to the desired comparability sought across tests.

Relative instability of item parameters

When examining how different the estimates of item easiness based on contrasting samples are, one should not expect the kind of closeness obtained with ability estimates. The comparison made on the level of item parameter estimates is more sensitive. It is also mostly relevant when the items are going to be combined into test forms of different composition than the one actually used in the calibration. According to Wright (1969), the best way to examine this question is to plot the contrasting calibrations for each item against each other and to fit a straight line to this plot with a slope of one but not necessarily with an intercept of zero. This is because the unstable items have a random translation effect on the good ones. The expected slope remains one but the intercept is moved away from zero by the bad items.

We followed the same procedure here as the one adopted for ability estimates. Appendices B1, B2 and B3 show the log easiness estimates for the eight sub-groups described in table 38. The first column contains the item numbers. The second column presents the estimates obtained from the total sample and the following eight columns show the estimates provided by the sub-samples. The standard errors of these estimates are presented in Appendices B4, B5, and B6. Again, if the hypothesis of no divergence holds true, the sum of the weighted squared difference between each log easiness estimate and a weighted average of all log easiness estimates should be zero for each item. We denoted this statistic as χ^2_n. The results are presented in tables 46,47, and 48. At a probability level of .001 and with 7 degrees of freedom, the critical chi-square value is 24.322. The number of items that shows fit on the basis of this statistic is 19 (38%) for the Quantitative Ability subtest, 14 (19%) for the Verbal Ability subtest, and 31 (36%) for the Science subtest. When comparing these results with the data of the preceding section, we conclude that the percentage of conformable items does not need to be high (20%) for the model to be adequate in terms of person measurement since no single score group shows any sign of misfit. On the other hand, if our purpose is to build up an item pool, the model provides enough information for item improvement and final selection.

The procedure was repeated for the sub-samples composed of subjects of different racial origins. Appendices B7, B8, and B9 show the log easiness estimates for the three sub-groups of table 42. The standard errors of these estimates are presented in Appendices B10, B11, and B12. The chi-square statistics are shown in tables 49, 50, and

TABLE 46

QUANTITATIVE ABILITY - DEGREE OF DIVERGENCE BETWEEN LOG EASINESS ESTIMATES COMPUTED FROM SUB-GROUPS OF TABLE 38

ITEM	$\hat{\theta}_o$	χ^2_n	ITEM	$\hat{\theta}_o$	χ^2_n
1	1.864	120.815	26	- .324	142.155
2	2.839	67.765	27	- .067	29.313
3	1.939	111.407	28	.121	12.999*
4	.578	39.157	29	- .955	15.221*
5	2.299	11.485*	30	- .835	95.213
6	1.553	42.213	31	- .513	12.526*
7	2.052	8.865*	32	-1.307	6.758*
8	1.313	7.340*	33	- .240	44.009
9	1.644	3.755*	34	.356	178.567
10	1.326	95.279	35	- .590	28.310
11	1.037	21.749*	36	- .043	21.883*
12	.815	95.998	37	- .891	79.502
13	.822	63.504	38	- .755	18.815*
14	.570	139.484	39	-1.403	20.822*
15	.165	31.471	40	- .639	55.203
16	.628	8.337*	41	-1.176	42.461
17	- .633	24.131*	42	- .949	4.907*
18	.468	4.186*	43	-1.374	203.731
19	.264	69.541	44	- .539	107.682
20	- .114	32.690	45	-1.032	32.171
21	- .278	13.142*	46	-1.435	247.358
22	- .996	287.603	47	-1.266	8.500*
23	.982	24.416	48	-1.416	67.016
24	- .347	11.918*	49	-2.155	55.096
25	- .104	251.111	50	-2.439	677.954

Key: $\hat{\theta}_o$: weighted average of log easiness estimates

χ^2_n : within-item chi-square with 7 degrees of freedom (critical value: 24.322)

* : Fitting items: 19 (38%)

TABLE 47

VERBAL ABILITY - DEGREE OF DIVERGENCE BETWEEN LOG EASINESS ESTIMATES COMPUTED FROM SUB-GROUPS OF TABLE 38

ITEM	$\hat{\theta}_o$	χ^2_n	ITEM	$\hat{\theta}_o$	χ^2_n
1	3.804	45.080	39	.039	21.303*
2	1.506	10.942*	40	- .332	174.839
3	1.414	109.293	41	- .030	192.702
4	1.073	414.268	42	- .111	297.139
5	1.212	239.685	43	- .257	5.828*
6	1.246	192.765	44	- .526	250.951
7	1.314	94.306	45	- .485	61.924
8	.701	5.236*	46	- .387	60.256
9	.825	85.276	47	-1.262	328.150
10	- .348	53.712	48	- .708	283.494
11	.753	11.440*	49	-1.177	70.564
12	.296	247.053	50	- .906	141.268
13	.041	156.655	51	-1.111	51.888
14	- .326	82.641	52	-1.314	18.647*
15	- .476	50.762	53	-1.136	204.034
16	- .511	21.677*	54	-1.206	477.928
17	- .321	22.854*	55	-1.609	156.066
18	- .202	112.916	56	1.762	145.583
19	- .619	7.140*	57	1.416	86.353
20	- .741	34.078	58	.143	777.936
21	- .785	27.856	59	.083	572.158
22	-1.230	123.275	60	.610	200.234
23	- .673	27.175	61	.470	21.122*
24	- .595	92.051	62	.122	34.943
25	- .815	41.638	63	.546	73.926
26	-1.106	10.907*	64	- .071	428.974
27	-1.224	22.027*	65	- .216	311.653
28	- .689	186.592	66	- .388	177.640
29	-1.764	87.210	67	- .198	490.252
30	-1.403	31.569	68	- .653	228.401
31	1.953	53.506	69	- .057	27.021
32	2.240	133.814	70	- .993	39.209
33	1.787	21.433*	71	- .510	245.069
34	.669	201.164	72	-1.618	230.138
35	.189	715.636	73	-1.227	4.085*
36	1.213	392.851	74	- .962	34.782
37	1.371	70.978	75	- .995	702.311
38	.767	34.616			

*: Fitting items: 14 (19%)

TABLE 48

SCIENCE - DEGREE OF DIVERGENCE BETWEEN LOG EASINESS ESTIMATES COMPUTED FROM SUB-GROUPS OF TABLE 38

ITEM	$\hat{\theta}_o$	χ^2_n	ITEM	$\hat{\theta}_o$	χ^2_n
1	1.653	10.542*	44	- .164	74.069
2	1.144	58.911	45	.020	3.352*
3	1.608	3.212*	46	- .302	22.524*
4	.812	33.595	47	- .146	7.403*
5	.153	92.219	48	- .415	15.691*
6	2.624	9.688*	49	- .328	95.905
7	.126	128.454	50	- .186	80.951
8	1.058	14.227*	51	.025	43.751
9	.492	168.244	52	- .468	11.351*
10	.153	129.773	53	- .260	10.424*
11	1.672	53.021	54	- .636	33.005
12	1.364	84.921	55	- .612	355.437
13	- .230	148.193	56	.999	145.629
14	- .666	274.236	57	-1.093	238.561
15	- .810	411.470	58	- .524	30.065
16	.809	108.952	59	- .525	334.723
17	.517	233.745	60	- .625	28.653
18	- .966	94.068	61	- .901	20.987*
19	1.305	32.789	62	- .540	127.027
20	1.316	9.606*	63	- .543	11.735*
21	1.244	191.970	64	-1.041	114.475
22	.656	22.022*	65	-1.095	35.009
23	.318	195.432	66	- .959	12.076*
24	.867	56.405	67	-1.659	58.610
25	.498	62.963	68	-1.405	11.534*
26	.891	132.231	69	-1.074	6.534*
27	.482	90.018	70	-1.615	83.326
28	.212	21.428*	71	-1.710	11.848*
29	.439	34.122	72	-1.925	12.330*
30	.341	22.387*	73	-1.559	18.643*
31	.146	67.795	74	.689	9.410*
32	.226	19.034*	75	.720	61.413
33	.135	85.513	76	- .105	13.883*
34	.547	159.456	77	.490	21.699*
35	- .079	181.368	78	.227	61.168
36	- .151	49.643	79	.195	85.449
37	.083	51.145	80	- .120	56.327
38	.210	97.229	81	-1.095	8.691*
39	.168	18.172*	82	.004	55.705
40	.715	79.405	83	- .865	58.338
41	- .068	7.394*	84	-1.336	243.469
42	- .132	21.328*	85	.204	18.477*
43	- .036	53.220	86	- .793	25.620

*Fitting items: 31(36%)

51. At a probability level of .001 and with 2 degrees of freedom, the critical chi-square value is 13.816. For this split, the number of fitting items is 36 (72%) for the Quantitative Ability subtest, 26 (35%) for the Verbal Ability subtest, and 49 (57%) for the Science subtest. These results shall be discussed later.

A question remains unanswered at this point. How can one know whether this new statistic is a more valid criterion of fit than the item chi-square described in Chapter II and used for the overall analyses in Chapter V? We shall examine this issue in the next section.

Test fit and item fit

The most obvious conclusion which comes out of our investigations is that one must differentiate between two notions of fit. Rentz and Bashaw (1975) shed some light on the issue of model-data fit. Using the conceptual framework proposed by these authors, we shall attempt to relate the results of Chapter VI to those of Chapter V.

In Chapter V, we examined the data provided by the total sample made up of 18,075 subjects within the context of a single overall analysis since this is what is generally done in practice. For each one of the three MCAT subtests, both overall tests of fit (chi-square and likelihood ratio) were not significant. We then stated that these results were not surprising given the large sample size used for the investigations. We can now show that for most practical situations these overall statistical tests are useless. To make this point clearer, we grouped in tables 52, 53, and 54 the results of the overall analyses together with the results of the eleven sub-analyses discussed

TABLE 49

QUANTITATIVE ABILITY - DEGREE OF DIVERGENCE BETWEEN LOG EASINESS ESTIMATES COMPUTED FROM SUB-SAMPLES OF TABLE 42

ITEM	$\hat{\theta}_o$	χ^2_n	ITEM	$\hat{\theta}_o$	χ^2_n
1	1.870	3.505*	26	- .290	12.534*
2	2.857	5.610*	27	- .070	22.941
3	1.954	7.224*	28	.147	12.552*
4	.590	4.230*	29	- .901	13.702*
5	2.247	17.140	30	- .801	14.587
6	1.519	16.253	31	- .481	1.906*
7	2.066	9.266*	32	-1.291	9.779*
8	1.303	1.973*	33	- .219	3.609*
9	1.647	13.475*	34	.393	11.655*
10	1.351	22.655	35	- .577	8.701*
11	1.013	9.392*	36	- .040	4.581*
12	.842	3.164*	37	- .855	1.526*
13	.838	32.649	38	- .750	1.577*
14	.632	16.684	39	-1.368	6.197*
15	.216	12.962*	40	- .637	7.400*
16	.648	11.434*	41	-1.141	.051*
17	- .620	45.769	42	- .943	8.088*
18	.484	.664*	43	-1.334	1.937*
19	.282	9.550*	44	- .492	2.436*
20	- .120	20.782	45	- .998	19.051
21	- .254	2.835*	46	-1.401	22.809
22	- .976	25.230	47	-1.228	2.460*
23	.964	7.855*	48	-1.381	2.496*
24	- .323	2.416*	49	-2.105	.120*
25	- .066	25.356	50	-2.452	102.616

Key: χ^2_n : within-item chi-square with 2 degrees of freedom (critical value:13.816)

*: Fitting items: 36(72%)

TABLE 50

VERBAL ABILITY - DEGREE OF DIVERGENCE BETWEEN LOG EASINESS ESTIMATES COMPUTED FROM SUB-GROUPS OF TABLE 42

ITEM	$\hat{\theta}_o$	χ^2_n	ITEM	$\hat{\theta}_o$	χ^2_n
1	3.642	100.226	39	.102	2.361*
2	1.502	56.217	40	- .271	.354*
3	1.429	112.878	41	.023	1.006*
4	1.194	108.035	42	- .055	15.320
5	1.305	93.424	43	- .213	3.590*
6	1.308	68.432	44	- .469	72.660
7	1.353	7.528*	45	- .438	21.918
8	.724	13.790*	46	- .341	9.239*
9	.846	5.474*	47	-1.209	24.621
10	- .320	.949*	48	- .681	31.673
11	.797	10.140*	49	-1.117	14.583
12	.344	36.027	50	- .848	49.704
13	.072	2.006*	51	-1.044	15.667
14	- .291	2.893*	52	-1.237	47.795
15	- .438	1.373*	53	-1.097	19.463
16	- .471	4.277*	54	-1.172	82.072
17	- .286	29.149	55	-1.567	26.100
18	- .165	29.553	56	1.705	241.500
19	- .581	12.564*	57	1.426	67.849
20	- .690	46.822	58	.216	12.715*
21	- .720	61.866	59	.284	40.394
22	-1.154	32.157	60	.680	44.251
23	- .620	6.067*	61	.511	3.683*
24	- .570	34.041	62	.143	2.959*
25	- .769	22.877	63	.557	18.313
26	-1.034	38.650	64	.012	17.510
27	-1.157	15.582	65	- .176	17.926
28	- .647	35.219	66	- .342	18.194
29	-1.682	11.221*	67	- .134	58.326
30	-1.324	10.362*	68	- .612	38.960
31	1.975	30.968	69	- .013	146.875
32	2.233	129.222	70	- .936	29.561
33	1.785	29.387	71	- .465	5.401*
34	.697	9.945*	72	-1.562	8.300*
35	.309	16.421	73	-1.143	3.883*
36	1.347	50.285	74	- .885	24.067
37	1.368	91.672	75	- .962	33.507
38	.790	1.455*			

*: Fitting items: 26(35%)

TABLE 51

SCIENCE - DEGREE OF DIVERGENCE BETWEEN LOG EASINESS ESTIMATES COMPUTED FROM SUB-GROUPS OF TABLE 42

ITEM	$\hat{\theta}_o$	χ^2_n	ITEM	$\hat{\theta}_o$	χ^2_n
1	1.659	3.121*	44	- .158	9.702*
2	1.164	10.541*	45	.017	17.617
3	1.604	.584*	46	- .300	10.744*
4	.826	3.237*	47	- .141	5.700*
5	.166	17.460	48	- .408	2.409*
6	2.608	20.480	49	- .305	2.835*
7	.151	42.140	50	- .164	1.348*
8	1.058	6.141*	51	.039	22.701
9	.509	27.906	52	- .455	.514*
10	.159	35.429	53	- .255	3.322*
11	1.663	22.646	54	- .636	25.166
12	1.370	30.163	55	- .607	59.884
13	- .198	13.404*	56	1.020	15.949
14	- .651	48.834	57	-1.072	57.157
15	- .819	81.099	58	- .517	8.171*
16	.815	10.933*	59	- .517	4.142*
17	.555	2.877*	60	- .621	16.068
18	- .974	4.080*	61	- .868	.545*
19	1.334	1.175*	62	- .507	3.280*
20	1.328	5.775*	63	- .518	3.547*
21	1.254	110.274	64	-1.041	8.787*
22	.644	22.566	65	-1.076	28.520
23	.327	15.648	66	- .944	13.870
24	.856	11.334*	67	-1.629	6.861*
25	.513	19.421	68	-1.387	10.939*
26	.895	2.506*	69	-1.077	1.296*
27	.474	3.407*	70	-1.611	26.116
28	.219	9.239*	71	-1.699	5.839*
29	.471	20.381	72	-1.916	3.416*
30	.327	3.745*	73	-1.533	35.859
31	.158	10.639*	74	.693	8.439*
32	.229	.883*	75	.719	13.644*
33	.157	4.709*	76	- .093	1.645*
34	.532	67.281	77	.479	46.475
35	- .060	63.717	78	.254	23.707
36	- .143	8.810*	79	.222	10.343*
37	.104	5.668*	80	- .118	14.814
38	.223	21.356	81	-1.084	2.542*
39	.185	13.003*	82	.007	42.745
40	.696	35.477	83	- .868	14.096
41	- .061	13.831	84	-1.314	19.355
42	- .104	9.642*	85	.243	41.573
43	- .029	4.592*	86	- .763	6.895*

* Fitting items: 49(57%)

earlier. For these sub-analyses, all but two of the overall chi-square tests of fit show misfit. The two exceptions are Verbal AMI where there are 74 items out of a total of 75 which fit and Science AMI where all of the 86 items fit. In these two cases however the likelihood ratio tests indicate misfit. As for the other sub-analyses, it is rather odd that with a very high proportion of fitting items (49 out of 50 in Quantitative AMI) both tests of fit show misfit. On the basis of such results, we conclude that these two overall tests of fit appear to be more misleading than useful and that, when dealing with large sample sizes, they should simply be ignored. Instead of calling them tests of fit, one could refer to them as tests of "perfection". This probably explains why so many potential users have rejected the Rasch approach to measurement claiming that the model did not fit most real testing situations. One can surely not conclude that the simple logistic model does not fit on the basis of such grounds alone.

Let us now turn our attention to the thesis developed by Rentz and Bashaw (1975), which we endorse totally. There exist two rather fundamentally different types of applications of the Rasch model that call for correspondingly different concepts of model-data fit. Rentz and Bashaw call the two types of applications test construction and test analysis and the corresponding concepts of fit, item fit for the former situation and test fit for the latter. The kind of freedom one has to manipulate the test at the item level is what constitutes the difference between the two situations. In the first type of application, that is, test construction, the test maker has the freedom to discard poor items and retain good ones. For this application, what is needed are indicators of fit for items. In the second type of application, that is, test

TABLE 52

QUANTITATIVE ABILITY - OVERALL TESTS OF FIT

SAMPLE	SIZE	OVERALL CHI-SQUARE		LIKELIHOOD RATIO		ITEM CHI-SQUARE	
		MEAN SQUARE	p	MEAN SQUARE	p	$p \geq .001$ FITTING	$p < .001$ NON FITTING
OVERALL	18,075	5.308	.000	1.129	.000	10	40
AMI	337	1.253	.000	3.789	.000	49	1
AMII	1,045	1.356	.000	4.597	.000	46	4
AMIII	3,111	1.948	.000	6.754	.000	39	11
AMIV	2,613	1.759	.000	9.728	.000	41	9
BMI	693	1.249	.000	-2.493	.000	46	4
BMII	1,575	1.704	.000	-8.208	.000	35	15
BMIII	3,206	1.864	.000	-8.873	.000	33	17
BMIV	2,187	1.693	.000	3.370	.000	39	11
W	10,685	3.867	.000	3.779	.000	14	36
B	626	1.185	.000	-1.011	.000	46	4
O	1,288	1.381	.000	0.022	.000	44	6

Key: AMI: Above median, income level I

. . .

. . .

. . .

BMIV: Below median, income level IV

W: White
B: Black
O: Other

TABLE 53

VERBAL ABILITY - OVERALL TESTS OF FIT

SAMPLE	SIZE	OVERALL CHI-SQUARE		LIKELIHOOD RATIO		ITEM CHI-SQUARE	
		MEAN SQUARE	p	MEAN SQUARE	p	$p \geq .001$ FITTING	$p < .001$ NON FITTING
OVERALL	18,075	6.156	.000	4.187	.000	6	69
AMI	374	1.043	.084	-4.076	.000	74	1
AMII	1,200	1.273	.000	3.875	.000	70	5
AMIII	3,221	1.911	.000	7.393	.000	56	19
AMIV	2,854	1.831	.000	5.488	.000	56	19
BMI	656	1.329	.000	-0.068	.000	66	9
BMII	1,420	1.506	.000	-1.312	.000	62	13
BMIII	3,096	2.112	.000	-5.211	.000	42	33
BMIV	1,946	1.624	.000	-2.917	.000	57	18
W	10,685	3.996	.000	5.635	.000	11	64
B	626	1.202	.000	-0.161	.000	66	9
O	1,288	1.470	.000	-2.561	.000	53	22

TABLE 54

SCIENCE - OVERALL TESTS OF FIT

SAMPLE	SIZE	OVERALL CHI-SQUARE		LIKELIHOOD RATIO		ITEM CHI-SQUARE	
		MEAN SQUARE	p	MEAN SQUARE	p	p≥.001 FITTING	p<.001 NON FITTING
OVERALL	18,075	2.996	.000	9.293	.000	27	59
AMI	325	0.996	.547	-3.956	.000	86	0
AMII	1,107	2.957	.000	1.247	.000	76	10
AMIII	3,203	1.348	.000	7.516	.000	78	8
AMIV	2,648	1.298	.000	2.238	.000	79	7
BMI	705	1.142	.000	-0.876	.000	82	4
BMII	1,513	1.228	.000	-2.859	.000	82	4
BMIII	3,114	1.292	.000	-6.040	.000	79	7
BMIV	2,152	1.287	.000	-5.169	.000	74	12
W	10,685	2.274	.000	4.750	.000	33	53
B	626	1.103	.000	-0.143	.000	83	3
O	1,288	1.175	.000	-2.808	.000	81	5

analysis, there is no freedom to discard poor items, the particular collection of test items being fixed. For this application, some measures of fit at the overall test level are required. In our work, we called the first situation test calibration instead of test construction and the second type of application person measurement instead of test analysis, thus referring to two normally consecutive stages in the process of measurement. The procedure for the evaluation of model-data fit proposed by Rentz and Bashaw (1975) rests on three sets of concepts: the assumptions of the model, some antecedent conditions and some consequent conditions. The three assumptions of the Rasch model, according to these authors, are the postulated logistic function translated into a probabilistic statement, the multiplicative rule pertaining to one item parameter and one person parameter, and the stochastic independence of all answers to a given test. In Chapter II of our study, only the stochastic independence was referred to as an assumption. The other two notions were considered simply as characteristics of the Rasch model. On the other hand, we identified as assumptions the unidimensionality of the trait being measured, equal item discriminations, and the absence of guessing. For Rentz and Bashaw, these are antecedent conditions which can be easily deduced from the assumptions but they are not assumptions. Similarly, Rentz and Bashaw argue that there is only one consequence component in Rasch's model, that is, specific objectivity, from which certain conditions may be deduced. These consequent conditions are essentially the stability or invariance of item parameters and ability parameters. Within such a framework, antecedent conditions are most likely to lead to indicators of item fit whereas consequent conditions might be most useful in describing test fit. Hence, Rentz and Bashaw

suggest the following definitions. Item fit is the extent to which items can be characterized according to those antecedent conditions derived from the model's assumptions. Test fit can be defined as the extent to which the test achieves those consequences specifiable from the concept of specific objectivity or as the extent to which the test contains fitting items in terms of a proportion of items that fit the model, using some specified criterion of item fit.

Obviously, our Chapter V dealt with antecedent conditions whereas the first sections of the present chapter were more concerned with consequent conditions. We shall now consider a few conclusions concerning test fit which will then be related to some of the item indicators described in Chapter V.

Indicators of model-data fit at the test level

For many practical situations, the stability of the ability parameter estimates is the most relevant aspect of model-data fit. Such a stability is a specific consequent condition of the Rasch model. The ability parameters of the model are supposed to be invariant with respect to any other person parameters. This means that any systematic variation in calibration conditions is inconsequential as long as scoring tables remain invariant. In order to determine the degree of invariance of these estimates, we computed a chi-square statistic for each score group in each MCAT subtest. Our first series of investigations comprised eight sub-groups. Each sub-group was homogeneous with respect to ability and parents' combined income so that there was much heterogeneity across sub-groups. Tables 39,40, and

41 showed that all scoring tables were totally invariant despite variations in sample sizes. Our second series of analyses considered three sub-groups composed of subjects of different racial origins. The data presented in tables 43, 44, and 45 led us to the same conclusions. Could such a chi-square statistic be used as a statistical test of invariance? Before answering this question, let us consider the stability index developed by Rentz and Bashaw. Since stability implies that a set of estimates of the same parameter will be invariant over repeated observations, Rentz and Bashaw used the standard deviation of the distribution of estimates as a measure of stability. Their stability index is simply the average of the standard deviations computed for a given sample of subjects at all score group levels. In order to assess the expected variability of the ability parameter estimates as a function of sample size, they drew 15 random samples from a total pool of 33,123 subjects (Vocabulary test) for each of four sample sizes: 500,1000,2000,4000. The test was made up of 30 items (29 score groups). They concluded that there was some tendency for the stability of the ability estimates to get better with increases in the size of the calibrating sample. However, this variation was quite small (from .022 to .017). The problem with such an index lies in its interpretation. We know that the ideal situation would be an index of zero. However, we do not know what the interval of acceptable values might be. Moreover, the standard deviation of a distribution of estimates does not take into account the error involved in the estimation itself. Our chi-square statistic obviates these difficulties. It gets essentially at the same thing as the stability index except that the average deviation of the estimates from their mean is weighted by the

inverse of the standard errors of these estimates. Furthermore, we found an almost perfect correlation between the stability indexes and our chi-square values (both computed from our data), so that the chi-square statistic seems to retain the qualities attributed by Rentz and Bashaw to the stability index. The main advantage of such an index is, according to Rentz and Bashaw, to allow for meaningful comparisons across different tests and across different analyses of the same test. The values of the chi-square statistic used in our analyses also possess this property.

On the basis of the fact that the three MCAT subtests have shown stability or invariance of the ability estimates at the level of all score groups, we conclude that the Rasch model fits the MCAT data for those applications where this kind of stability is most relevant. Such applications include measurement proper, linking and equating of different test forms. Interestingly, our results with respect to race correspond to what was found by Rentz and Bashaw with the same levels of that variable, that is, "practically identical ability parameter estimates." Here is their conclusion:

> "Whenever studies like this are conducted, where stability is observed across samples differing in composition, the variable most closely related to the latent trait being measured by the test will show the greatest instability, as long as the test contains items with less than perfect model-data fit." (Rentz and Bashaw, 1975).

Another indicator of model-data fit at the test level is the invariance of item easiness estimates. There are applications where item stability would be significantly more important than the stability of ability estimates. One such application is tailored testing (Rentz and Bashaw, 1975). In this situation, there is an attempt to match items to subjects, "and since misfit affects the stability of items more than

it does the abilities, a higher degree of fit would be required than that necessary for applications requiring only stable ability estimates".

Rentz and Bashaw used the same stability index here as the one applied to ability parameter estimates. They found however that easiness estimates were more sensitive to different sample sizes than were ability estimates. They explained this phenomenon as follows:

> "The basic observation for estimating easiness is, p, the proportion answering the item correct, a number whose accuracy depends directly on the sample size. The stability of ability estimates depends on both the item easiness estimates and the number of items. Thus the extent of sample size influence on ability estimates is limited by its influence on item easiness. Furthermore the influence of item easiness variability tends to attenuate as the number of items become greater. The consequence of the interplay of these factors is the observed difference between the stability of the easiness and ability estimates." (Rentz and Bashaw, 1975).

Using the chi-square statistic previously applied to scoring tables, we tested the stability of item parameter estimates and found invariance in 19 items (38%), 14 items (19%), and 31 items (36%), for the Quantitative Ability subtest, the Verbal Ability subtest, and the Science subtest respectively, in the first series of investigations (Tables 46, 47, and 48). The second series of investigations gave 36 items (72%), 26 items (35%), and 49 items (57) showing invariance for the same subtests (Tables 49, 50, and 51). According to these figures, the best fitting test would be the Quantitative Ability subtest which is the one containing the smallest number of items (50). The second best fitting test would be the Science subtest with 86 items and the worst fitting test would be the Verbal Ability subtest which is composed of 75 items. The only firm conclusion we could make on the basis of those results was that a set

of "conformable" items as low as 20% in a given test seemed sufficient to ensure stability at the ability estimates level. This result was obtained with a 75 item test though. Considering the remarks made by Rentz and Bashaw about the effect of the number of items on the stability of ability estimates, other empirical studies need be conducted before such generalizations could be made.

Hence, measurement is possible with less than perfect tests. For the kinds of applications where item stability is required, the most logical attitude is to work with stable items. As for unstable items, one must find out the most likely reasons that could explain their behaviors. The main advantage of our chi-square test of invariance is that of diagnosing those items suffering from imperfection. To detect the actual degree of imperfection and its causes, we must resort to indicators of fit at the item level.

Indicators of model-data fit at the item level

Among the antecedent conditions necessary for the Rasch model to represent data adequately is the unidimensionality of the trait being measured. In the present study, we did not explore this aspect of model-data fit. However, Rentz and Bashaw (1975) used an index of first factor concentration which was derived from a principal components analysis of the item intercorrelation matrices of their 14 tests. The index represents the percentage of variance accounted for by the first component. For the 14 tests analyzed, the index varied from 17.4% to 30.7%

The most general index of fit of items to the model is the magnitude of the mean squares. We have already insisted on the overly sensitive nature of such an indicator of fit. Because this index is a function of sample size, its interpretation must be made with care.

> "The problem of interpreting the mean squares is a general problem in statistical hypothesis testing. The role of large samples in rejecting null hypotheses is well known. For any difference between data and an hypothesis, most statistical tests will lead to the rejection of the null hypothesis if the sample is large enough". (Rentz and Bashaw, 1975).

The mean square fits are based on the difference between expected and obtained proportions for each item-by-score group cell entry. With a fairly large sample size, this difference is estimated quite accurately. But if the sample size is 10 times larger, the value of the mean square will increase ten-fold. Rentz and Bashaw argued that the item mean squares would be a defensible choice as an index of test fit since any factor that might cause misfit would be reflected in the mean squares. To control for large sample sizes, they suggested the use of a correcting factor applied to the mean or the median of all item mean squares in a test. This factor was 10,500/N for their data, 10,500 being the smallest sample size they had in their analyses. For their 14 tests, Rentz and Bashaw found average mean squares (adjusted for large sample sizes) ranging from 5.4 to 10.9. They compared those values with that obtained by Cartledge (1974) on simulated data. Using the same correcting factor, Cartledge found average mean squares of about 2.0. Rentz and Bashaw concluded that their tests were neither very good nor very bad fitting tests. If we compare our data and results to those of Rentz and Bashaw and those of Cartledge, we get the following figures for the three MCAT subtests (overall runs):

	Quantitative	Verbal	Science
Unadjusted average mean squares	5.308	6.156	2.996
Adjusted(10,500/18,075) average mean squares	3.083	3.576	1.740

The median values are even better: 3.8, 4.6, and 2.3 (unadjusted); 2.2, 2.6, and 1.3 (adjusted) for the Quantitative, Verbal and Science subtests in that order. We must conclude that the three MCAT subtests show reasonable fit on this overall criterion. This time, it is the Science subtest that seems to be the best fitting test, the Verbal subtest still being last. As can be seen in tables 52, 53, and 54, there is a very high correlation between the size of the sample, the magnitude of the average mean square, and the proportion of fitting items (item mean square fit). The smaller the sample size, the lower the magnitude of the average mean square in a test and the larger the proportion of fitting items. It seems as if perfect fit is obtained with a sample size of about 300 (AMI in tables 52, 53, 54). The sample size effect is therefore difficult to disentangle from real misfit conditions for samples of size greater than 300. Further studies are needed to elucidate this question.

Another important indicator of model-data fit at the item level is an index of item discrimination. Equal item discrimination constitutes one of the antecedent conditions required for the consequent condition of specific objectivity to manifest itself. We discussed this issue at length in Chapter V. The criterion used was the slope, that is, the regression of item log odds on test log odds. Rentz and Bashaw (1975) determined the relative number of items for which the slope fell in the

interval 1.0 plus or minus .2 for each of their 14 tests and used that proportion as a slope index of fit. Their values ran from 46.6% to 65.0%. For each MCAT subtest this proportion was 30% (Table 15). Rentz and Bashaw also used the semi-interquartile range as a measure of dispersion of the distribution of item slopes. We computed the standardized difference of the slope from unity (adjusted for small standard errors) instead of the semi-interquartile range. The proportion of fitting items on this criterion was 62% for the Quantitative, 44% for the Verbal, and 45% for the Science subtest (Tables 35, 36, 37).

Finally, we computed a correlation coefficient between normal deviates and score groups to obtain another indicator of test fit at the item level that would get at a possible guessing effect or speed factor. The relative number of fitting items here was 56% for the Quantitative, 33% for the Verbal, and 45% for the Science subtest (Tables 35,36,37).

In summary, the relationship between consequent and antecedent conditions is presented in table 55 where indicators of fit at the test level can be compared with indicators of fit at the item level. A formal link is thus established between Chapter V and Chapter VI.

TABLE 55

DEGREE OF FIT OF THE THREE MCAT SUBTESTS

SUBTEST	CONSEQUENT CONDITIONS			
	ABILITY INVARIANCE*		ITEM INVARIANCE**	
	FIRST SPLIT (J=8)	SECOND SPLIT (J=3)	FIRST SPLIT (J=8)	SECOND SPLIT (J=3)
QUANTITATIVE (K=50)	100	100	38	72
VERBAL (K=75)	100	100	19	35
SCIENCE (K=86)	100	100	36	57

KEY J: number of sub-groups
K: number of items
*: each cell entry = percentage of score groups showing fit
**: each cell entry = percentage of items showing fit

SUBTEST	ANTECEDENT CONDITIONS				
	MEAN SQUARE		ITEM DISCRIMINATION		GUESSING/SPEED
	MEAN*	MEDIAN*	SLOPE** (1.0± .2)	SDU** (<3)	ND-SG** (r)
QUANTITATIVE	3.083	2.2	30	62	56
VERBAL	3.576	2.6	30	44	33
SCIENCE	1.740	1.3	30	45	45

KEY *: adjusted for large sample size (10,500/18,075)
**: percentage of items showing fit on that criterion
SDU: standardized difference of slope from unity (adjusted for small standard errors)
ND-SG: correlation coefficient between normal deviates and score groups

Convergence of results

In Chapter V, we concluded that the best fitting items would be the items which would show fit on the three criteria examined, that is, item mean square, standardized difference of slope from unity, and correlation between normal deviates and score groups. As a corollary, we added that the worst fitting items would be those which would show misfit on the same criteria. Since these criteria were used to evaluate whether or not the antecedent conditions required by the Rasch model were respected, there should be a close relationship between the conclusions reached at that level and the conclusions one would draw in examining the stability of item easiness estimates. Such a relationship is essential for item selection.

It turns out that there is an almost perfect correlation in the identification of best and worst fitting items between the three criteria taken together and χ^2_n used as a test of item invariance. Table 56 shows that χ^2_n failed to identify only one item in the three subtests which was identified as a good item by the three criteria. Moreover, only three items out of the 64 which were categorized as misfitting on the three criteria showed fit on χ^2_n.

Such a finding tends to validate the procedure suggested in the last section of Chapter V for examining different potential sources of misfit. It also establishes a link between indicators of fit at the item and test levels which is more explicit than the one presented in table 55.

TABLE 56

BEST AND WORST FITTING ITEMS

SUBTEST	ANTECEDENT CONDITIONS FIT ON 3 CRITERIA	CONSEQUENT CONDITION FIT ON χ^2_n
	ITEMS	ITEMS
Quantitative	9,18,21,24,27,31,42 (n=7)	9,18,21,24,31,42 (n=6)
Verbal	8,43 (n=2)	8,43 (n=2)
Science	1,3,6,20,22,41,61,69,72, 74,76,85 (n=12)	1,3,6,20,22,41,61,69, 72,74,76,85 (n=12)
Total	21	20
	MISFIT ON 3 CRITERIA	MISFIT ON χ^2_n
	ITEMS	ITEMS
Quantitative	1,12,14,22,26,28,30,35, 41,46 (n=10)	1,12,14,22,26,30,35,41, 46 (n=9)
Verbal	4,10,22,24,28,29,35,36, 40,42,44,47,48,50,53,54, 55,58,59,60,64,65,66,67, 68,71,72,74,75 (n=29)	4,10,22,24,28,29,35,36, 40,42,44,47,48,50,53,54, 55,58,59.60,64,65,66,67, 68,71,72,74,75 (n=29)
Science	4,5,7,13,14,15,16,27,28, 33,35,36,37,49,52,55,56, 57,64,70,78,79,80,83,84 (n=25)	4,5,7,13,14,15,16,27,33, 35,36,37,49,55,56,57,64, 70,78,79,80,83,84 (n=23)
Total	64	61

CHAPTER VII

CONCLUSIONS AND IMPLICATIONS

Since 1970 the number of papers on the Rasch model has increased tremendously. Yet, the problem of fit has not been satisfactorily solved. An important problem this study has addressed itself to is the investigation of the robustness of the simple logistic model to some violations of its assumptions. This question was examined at length on simulated data by Nargis Panchapakesan. However, as she pointed out, her results needed to be applied to real data in order to be considered valid. One major difference between simulated and real data lies within some "external" criteria which are required to validate the procedures used for the analysis. With simulations, one controls the true state of affairs. For instance, the difference between generating parameters and estimated parameters can be used as an external criterion against which the effect of some departure of the data from the specifications of the model can be assessed. In Panchapakesan's work only one departure at a time was explored so that the question of knowing whether or not different sources of misfit confound one another remained unanswered.

Our conclusions can be grouped into two categories. The first set concerns the test itself. What can we say about the MCAT test on the basis of our investigations with the Rasch model? The second set is related to basic issues in the area of model-data fit. How can fit be defined? Are there different types of fit for different kinds of applications? Can various causes of misfit be disentangled, that is, identified separately? We shall examine these questions in summarizing

our results.

The MCAT test

The MCAT test is composed of four subtests. It is used by medical educators to make decisions about who is going to be accepted for medical training in this country. For this study, we retained the three most important subtests: quantitative, verbal, and science. The main concern of those who make decisions on the basis of the results obtained by examinees on these subtests has to do with the fairness of the MCAT to various subgroups within the population of examinees. One way of examining this issue is to split the overall sample into homogeneous subgroups on the variable of interest and to consider how divergent the scoring tables obtained from each subgroup would be. The advantage of the Rasch model over any other psychometric model is that of providing a sound mathematical basis for putting all scores on a common scale so that such a comparison is meaningful. Scoring tables simply establish a correspondence between raw scores on a test and a measure of ability on a natural log scale. If a test were such that a given raw score obtained by a group of examinees would yield a measure of ability different than the one obtained by another group of examinees, that test could be considered as unfair. The reverse statement is a necessary but not a sufficient condition for test fairness. In Chapter VI, we established that the scoring tables computed from eight subgroups in one case and from three subgroups in the other case were equivalent for the three MCAT subtests (Tables 39-41 and 43-45). This means that any of the three MCAT subtests could have been calibrated

on any of the eleven subgroups and applied to the other subgroups without fear of reaching different estimates of ability. These results illustrate two things. First, they indicate that estimates of ability are really free from sample considerations under the simple logistic model, and second, that there is no reason to worry about getting a proper standardizing sample for test calibration when using this model. We thus conclude that the Rasch model fits the MCAT data, that is, applies to its three aptitude subtests and is not influenced by three of its population's characteristics: intellectual ability, socio-economic status (parents' income level), and race. This is true at the measurement level at least. A test is never perfect. But on the basis of these results, one can say that no matter how imperfect the MCAT really is, this does not affect the measurement proper. The invariance of ability estimates is a necessary first degree in model-data fit. According to Rentz and Bashaw (1975) it is the single most important aspect of fit for one application area, namely equating of test forms. It is also the kind of fit which is the least sensitive to variations in sample sizes.

The MCAT test was submitted to an investigation of another aspect of fit, that provided by the degree of invariance of item easiness estimates computed from the eleven subgroups identified earlier. In the first case (eight subgroups), the percentage of items showing fit was 38 for the Quantitative Ability subtest, 19 for the Verbal Ability subtest, and 36 for the Science subtest. In the second case (3 subgroups), the percentage were respectively 72, 35, and 57 (Tables 46-51). How can these results be interpreted? If we relate those percentages to the 100% of stability found at the level of ability estimates, we conclude that a

percentage of "conformable" items as low as 19% is sufficient for the model to be adequate in terms of person measurement. However, a few questions remain unanswered. What is the minimum number of invariant or stable items in a test required to ensure stability at the level of ability estimates? It would be quite disturbing to find out that such a stability could be reached with a test composed entirely of unstable items. Further research is needed to elucidate this matter. Another issue that needs investigation is the effect of unstable items on good items. In this study, we did not delete any of the worst items and recalibrate the test. Had we done that, we might have discovered that the percentages of stable items would in fact have been higher than those reported. We must also consider that the sensitivity of item easiness estimates to variations in sample sizes is greater than that of ability estimates as was clearly shown by Rentz and Bashaw (1975). Such a variation is a confounding variable in our study. For instance, the sample sizes used for our racial split are 10,685 for Whites, 626 for Blacks, and 1,288 for Others. Interestingly enough, the proportion of misfitting items (using the indicators of item fit described in Chapter V) is directly proportional to the size of the sample for each MCAT subtest. In order to disentangle the effect of sample size from other likely hypotheses in explaining the instability of item as well as ability parameters, one needs to establish a base line. Further studies should take this effect into account by drawing random samples of equal size in each level of a variable of interest. But as was mentioned in Chapter VI, test fit can also be determined in terms of some proportion of items meeting some established criteria. We shall not repeat here the conclusions discussed at length in Chapter VI.

Yet, in judging the degree of fit of the MCAT, we must add the following considerations. Compared to the results obtained by Cartledge (1974) and by Rentz and Bashaw (1975), using an adjusted median mean square, the three MCAT subtests show a high degree of fit. As for an overall slope index of fit (the interval $1.0 \pm .2$ determined by the standardized difference of the item slope from unity) our results are grossly equivalent to those obtained by Rentz and Bashaw, the percentage of fitting items being 62 for the Quantitative, 44 for the Verbal, and 45 for the Science subtest. Here again we must add that any slope index of fit would be influenced by the size of the sample and that care must be exercised in its interpretation. Finally, using a correlation coefficient between normal deviates and score groups, we determined that the percentage of fitting items was 56 for the Quantitative, 33 for the Verbal, and 45 for the Science subtest.

We conclude that the three MCAT subtests show a perfect fit at the level of ability parameter estimates and a moderate degree of fit at the level of item invariance. These conclusions entail some practical implications for the MCAT use other than establishing the absence of major SES and race bias. The degree of fit found at the level of ability parameter estimates suggests that the Rasch model could be used for the equating of test forms, as shown by Rentz and Bashaw (1975). It also provides an incentive to initiate studies of the predictive validity type in which log ability estimates could be used instead of standardized scores as predictors. As for the degree of fit found at the level of item easiness estimates, there are many indications in this dissertation about possible explanations for this which might be useful for test construction and item selection. Only those responsible for the MCAT

program could draw conclusions that would make sense to them.

Model-data fit

The most important objective of this study was to establish a link between different indicators of fit at the item and test levels. We found that there was an almost perfect correlation between the three criteria of item fit presented in Chapter V (item mean square, standardized difference of slope from unity, and correlation between normal deviates and score groups) taken together and the chi-square of item invariance used in Chapter VI. We further stated that there was a logical relationship between the invariance of easiness estimates and ability estimates. On the basis of our study and the research already published, we summarized in table 57 the different kinds of fit that have been identified, the application areas for which each kind of fit is most relevant, and the criteria available for making fit-misfit decisions. We think that the sequence presented in table 57 should be the one followed in practice in exploring the fit of the Rasch model to some set of data.

We conclude that the criteria used by Panchapakesan can be applied to real data. We thoroughly examined the criterion of 1.0 ± 0.2 for the slope using her suggested index, that is, the standardized difference of the slope from unity. We reached the same conclusions. Items for which the slope is 0.4 or less are the worst items in a test no matter which external indicator of fit is taken for comparison. Similarly, items in the range 0.8 - 1.2 are the best items and there is a gray area in the range 0.4 - 0.8. As far as guessing and speed are

TABLE 57

DIFFERENT KINDS OF FIT FOR DIFFERENT APPLICATIONS

TYPE OF FIT	APPLICATION	CRITERIA
Consequent condition: Stability of ability parameter estimates	Measurement proper Equating of test forms	χ^2 test of invariance Stability index (Rentz and Bashaw, 1975)
Consequent condition: Stability of easiness parameter estimates	Self-tailored testing Sequential testing Test construction Item selection	χ^2 test of invariance Stability index (Rentz and Bashaw, 1975)
Antecedent conditions: All together (at the test level)	Test construction	Average mean square Median mean square
Antecedent conditions: All together (at the item level)	Item improvement (mis-scoring) Item selection	Item mean square (proportion of fitting items)
Antecedent condition: Unidimensionality (at the test level)	Measurement proper Test improvement	Factor analysis of inter-item tetrachoric correlations principal components analysis (first factor concentration)
Antecedent condition: Item discrimination (at the test level)	Test construction Item improvement (mis-scoring)	Interval 1.0 ± 0.2 Standardized $\bar{d}$ifference of slope from unity (proportion of fitting items)
Antecedent condition: Guessing (at the test level)	Test improvement	Average ability of sample greater than average difficulty of test. No subjects with a score lower than r* (Panchapakesan, 1969)

TYPE OF FIT	APPLICATION	CRITERIA
Antecedent condition: Guessing (at the item level)	Item improvement Test construction	Correlation between normal deviates and score groups (significant and negative) for difficult items
Antecedent condition: Speed (at the test level)	Test improvement	Percentage of subjects who did not finish the test
Antecedent condition: Speed (at the item level)	Test improvement	Percentage of subjects who omitted the answer Correlation between normal deviates and score groups (significant and positive) for last third of items in the test

concerned, we explored the potential value of a criterion suggested but not used by Panchapakesan, the correlation between normal deviates and score groups. In the last section of Chapter V, we analyzed the relationships existing between the two indicators of fit at the item level mentioned above and the item chi-square. We proposed a set of conclusions which were validated by the very fact that the three criteria taken together turned out to be almost perfect predictors of the invariance found at the level of item easinesses.

In this study, we tried very hard to relate different indicators of fit to one another. We believe that it is only when such a correspondence is clearly established that one could really disentangle the many possible sources of misfit in test construction. We think that this goal was only partially achieved here. However this study illustrated very neatly how the Rasch model can be used with confidence for scaling and measurement purposes.

APPENDICES

APPENDIX A

APPENDIX A 1

QUANTITATIVE ABILITY - SCORING TABLES, FIRST SPLIT

SCORE GROUP	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
1	-4.606	-4.685	-4.718	-4.583	-4.625	-4.525	-4.648	-4.686	-4.677
2	-3.849	-3.930	-3.958	-3.835	-3.874	-3.774	-3.882	-3.916	-3.909
3	-3.381	-3.463	-3.486	-3.375	-3.410	-3.311	-3.407	-3.438	-3.431
4	-3.031	-3.113	-3.132	-3.033	-3.064	-2.967	-3.051	-3.079	-3.073
5	-2.747	-2.829	-2.844	-2.754	-2.783	-2.688	-2.762	-2.788	-2.782
6	-2.505	-2.584	-2.596	-2.517	-2.542	-2.451	-2.516	-2.539	-2.534
7	-2.292	-2.368	-2.378	-2.307	-2.329	-2.242	-2.299	-2.320	-2.315
8	-2.100	-2.173	-2.180	-2.118	-2.138	-2.053	-2.103	-2.122	-2.118
9	-1.924	-1.994	-1.999	-1.944	-1.961	-1.881	-1.925	-1.942	-1.937
10	-1.761	-1.827	-1.830	-1.783	-1.798	-1.722	-1.759	-1.775	-1.770
11	-1.608	-1.670	-1.672	-1.631	-1.644	-1.572	-1.604	-1.618	-1.614
12	-1.463	-1.521	-1.522	-1.488	-1.498	-1.431	-1.458	-1.470	-1.466
13	-1.326	-1.379	-1.379	-1.351	-1.360	-1.297	-1.320	-1.330	-1.326
14	-1.195	-1.243	-1.243	-1.220	-1.227	-1.168	-1.187	-1.196	-1.192
15	-1.068	-1.113	-1.111	-1.094	-1.099	-1.045	-1.060	-1.067	-1.064
16	- .946	- .986	- .984	- .971	- .975	- .925	- .937	- .943	- .940
17	- .828	- .863	- .861	- .853	- .855	- .809	- .818	- .822	- .819
18	- .713	- .744	- .741	- .737	- .738	- .697	- .703	- .705	- .703
19	- .600	- .627	- .623	- .623	- .624	- .586	- .590	- .591	- .588
20	- .490	- .512	- .509	- .512	- .512	- .478	- .479	- .479	- .477
21	- .382	- .400	- .396	- .403	- .401	- .372	- .371	- .369	- .367
22	- .275	- .289	- .285	- .295	- .293	- .268	- .264	- .261	- .259
23	- .170	- .179	- .175	- .188	- .185	- .164	- .158	- .154	- .153
24	- .065	- .071	- .066	- .083	- .078	- .062	- .054	- .048	- .047
25	.039	.037	- .042	.023	.028	.040	.050	.057	.057
26	.143	.145	.149	.128	.133	.142	.154	.161	.162
27	.246	.252	.257	.233	.239	.243	.257	.265	.266
28	.350	.359	.364	.338	.345	.345	.361	.370	.370
29	.454	.467	.472	.443	.451	.447	.465	.475	.474
30	.559	.576	.581	.550	.558	.550	.569	.580	.579
31	.665	.685	.690	.658	.667	.654	.675	.687	.686
32	.773	.797	.802	.767	.777	.759	.782	.795	.793
33	.882	.910	.915	.879	.889	.866	.892	.905	.903
34	.994	1.025	1.030	.993	1.003	.976	1.003	1.017	1.014
35	1.109	1.144	1.149	1.110	1.121	1.088	1.117	1.131	1.129
36	1.228	1.266	1.271	1.231	1.242	1.203	1.235	1.250	1.246
37	1.350	1.393	1.398	1.356	1.368	1.323	1.357	1.372	1.368
38	1.478	1.525	1.530	1.487	1.499	1.448	1.484	1.499	1.495
39	1.613	1.663	1.668	1.624	1.637	1.579	1.618	1.633	1.628
40	1.755	1.810	1.815	1.769	1.782	1.717	1.758	1.774	1.769
41	1.906	1.966	1.971	1.924	1.938	1.865	1.909	1.924	1.918
42	2.070	2.135	2.140	2.090	2.106	2.024	2.071	2.086	2.080
43	2.249	2.320	2.325	2.275	2.291	2.198	2.248	2.262	2.256

APPENDIX A 1 - Continued

SCORE GROUP	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
44	2.448	2.525	2.531	2.479	2.496	2.392	2.445	2.459	2.452
45	2.676	2.760	2.766	2.713	2.730	2.614	2.670	2.684	2.676
46	2.944	3.036	3.043	2.988	3.007	2.876	2.935	2.949	2.941
47	3.277	3.379	3.386	3.329	3.350	3.202	3.264	3.277	3.269
48	3.729	3.840	3.849	3.790	3.814	3.645	3.711	3.723	3.714
49	4.468	4.592	4.603	4.541	4.569	4.376	4.445	4.457	4.447

Key: The entries represent the log ability estimates computed for each sub-group and for each score group.

OVERALL: log ability estimates computed from the total sample of subjects

AMI: above median, income level I
⋮ ⋮
BMIV: below median, income level IV

APPENDIX A 2

VERBAL ABILITY - SCORING TABLES, FIRST SPLIT

SCORE GROUP	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
1	-5.036	-5.076	-5.570	-5.627	-5.323	-4.734	-4.895	-5.008	-5.082
2	-4.262	-4.327	-4.680	-4.766	-4.550	-3.997	-4.136	-4.229	-4.290
3	-3.788	-3.868	-4.138	-4.231	-4.073	-3.552	-3.674	-3.754	-3.805
4	-3.442	-3.529	-3.748	-3.840	-3.719	-3.227	-3.336	-3.407	-3.451
5	-3.166	-3.256	-3.442	-3.528	-3.435	-2.969	-3.068	-3.131	-3.170
6	-2.936	-3.025	-3.188	-3.267	-3.194	-2.754	-2.843	-2.901	-2.936
7	-2.737	-2.823	-2.969	-3.041	-2.983	-2.568	-2.650	-2.702	-2.734
8	-2.561	-2.644	-2.776	-2.840	-2.796	-2.404	-2.479	-2.527	-2.556
9	-2.402	-2.481	-2.602	-2.660	-2.625	-2.256	-2.325	-2.370	-2.395
10	-2.257	-2.331	-2.443	-2.494	-2.468	-2.121	-2.185	-2.226	-2.249
11	-2.123	-2.192	-2.295	-2.341	-2.322	-1.997	-2.055	-2.093	-2.114
12	-1.999	-2.062	-2.158	-2.198	-2.185	-1.881	-1.935	-1.970	-1.989
13	-1.882	-1.939	-2.028	-2.063	-2.056	-1.773	-1.822	-1.855	-1.872
14	-1.771	-1.823	-1.906	-1.936	-1.934	-1.670	-1.716	-1.745	-1.761
15	-1.666	-1.712	-1.789	-1.815	-1.817	-1.573	-1.615	-1.642	-1.656
16	-1.566	-1.607	-1.678	-1.700	-1.705	-1.480	-1.518	-1.543	-1.555
17	-1.470	-1.505	-1.571	-1.589	-1.597	-1.392	-1.426	-1.449	-1.459
18	-1.378	-1.407	-1.468	-1.483	-1.494	-1.306	-1.337	-1.358	-1.367
19	-1.289	-1.313	-1.369	-1.381	-1.394	-1.224	-1.252	-1.270	-1.278
20	-1.203	-1.222	-1.273	-1.282	-1.297	-1.144	-1.169	-1.186	-1.192
21	-1.119	-1.133	-1.180	-1.186	-1.203	-1.067	-1.089	-1.104	-1.109
22	-1.038	-1.047	-1.089	-1.093	-1.111	- .992	-1.011	-1.024	-1.028
23	- .958	- .962	-1.001	-1.002	-1.022	- .919	- .935	- .946	- .949
24	- .881	- .880	- .915	- .914	- .935	- .847	- .861	- .870	- .872
25	- .805	- .800	- .831	- .828	- .850	- .777	- .788	- .796	- .797
26	- .731	- .721	- .749	- .744	- .767	- .708	- .717	- .723	- .723
27	- .658	- .644	- .668	- .662	- .685	- .641	- .647	- .651	- .651
28	- .587	- .568	- .589	- .581	- .605	- .574	- .578	- .581	- .579
29	- .516	- .494	- .511	- .502	- .527	- .509	- .510	- .511	- .509
30	- .446	- .420	- .434	- .424	- .449	- .444	- .443	- .443	- .440
31	- .378	- .347	- .359	- .347	- .373	- .380	- .377	- .376	- .372
32	- .310	- .276	- .285	- .272	- .298	- .317	- .312	- .309	- .304
33	- .242	- .205	- .211	- .197	- .223	- .254	- .247	- .243	- .237
34	- .176	- .134	- .138	- .124	- .150	- .191	- .183	- .177	- .171
35	- .109	- .065	- .066	- .051	- .077	- .129	- .119	- .112	- .105
36	- .043	.005	.005	.021	- .005	- .068	- .055	- .047	- .040
37	.022	.074	.076	.093	.066	- .006	.008	.018	.025
38	.088	.142	.146	.164	.137	.056	.072	.082	.090
39	.153	.211	.216	.235	.208	.117	.135	.147	.155
40	.218	.279	.286	.305	.278	.178	.198	.211	.220
41	.283	.347	.356	.376	.349	.240	.261	.275	.285
42	.349	.415	.426	.446	.419	.302	.325	.340	.350
43	.414	.484	.495	.516	.489	.364	.388	.404	.415
44	.480	.553	.565	.586	.559	.426	.452	.469	.480

APPENDIX A 2 - Continued

SCORE GROUP	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
45	.546	.622	.635	.657	.630	.489	.517	.535	.546
46	.613	.691	.705	.727	.701	.553	.582	.601	.612
47	.680	.761	.776	.799	.772	.617	.647	.667	.679
48	.748	.832	.848	.870	.844	.682	.713	.734	.747
49	.817	.904	.920	.943	.916	.748	.781	.803	.815
50	.887	.977	.993	1.016	.990	.814	.849	.872	.885
51	.958	1.051	1.067	1.090	1.064	.882	.918	.942	.955
52	1.030	1.126	1.142	1.165	1.139	.952	.989	1.013	1.027
53	1.103	1.203	1.218	1.242	1.216	1.023	1.061	1.086	1.101
54	1.178	1.282	1.296	1.320	1.294	1.095	1.134	1.161	1.175
55	1.255	1.362	1.375	1.400	1.374	1.169	1.210	1.237	1.252
56	1.334	1.445	1.457	1.482	1.456	1.246	1.288	1.316	1.331
57	1.416	1.531	1.541	1.566	1.541	1.325	1.368	1.396	1.412
58	1.500	1.619	1.628	1.652	1.627	1.406	1.451	1.480	1.496
59	1.587	1.711	1.717	1.742	1.717	1.491	1.537	1.567	1.583
60	1.678	1.807	1.810	1.835	1.811	1.579	1.626	1.657	1.674
61	1.772	1.908	1.907	1.932	1.908	1.672	1.720	1.751	1.768
62	1.872	2.015	2.009	2.034	2.010	1.769	1.818	1.851	1.868
63	1.977	2.128	2.117	2.142	2.118	1.872	1.922	1.955	1.973
64	2.089	2.249	2.231	2.255	2.232	1.982	2.033	2.067	2.085
65	2.209	2.380	2.353	2.377	2.354	2.099	2.152	2.186	2.204
66	2.339	2.522	2.485	2.509	2.486	2.227	2.280	2.316	2.334
67	2.480	2.681	2.628	2.653	2.630	2.366	2.421	2.457	2.475
68	2.638	2.859	2.788	2.812	2.790	2.522	2.577	2.614	2.633
69	2.815	3.066	2.967	2.991	2.969	2.697	2.754	2.791	2.810
70	3.021	3.313	3.175	3.199	3.177	2.901	2.958	2.996	3.016
71	3.267	3.624	3.423	3.446	3.425	3.145	3.203	3.242	3.262
72	3.577	4.053	3.735	3.758	3.738	3.453	3.513	3.552	3.572
73	4.005	4.770	4.165	4.188	4.168	3.879	3.940	3.980	4.000
74	4.721	4.975	4.883	4.905	4.886	4.592	4.654	4.695	4.715

APPENDIX A 3

SCIENCE - SCORING TABLES, FIRST SPLIT

SCORE GROUP	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
1	-4.834	-4.880	-4.913	-4.915	-4.938	-4.768	-4.808	-4.837	-4.830
2	-4.113	-4.162	-4.193	-4.193	-4.213	-4.050	-4.089	-4.116	-4.109
3	-3.681	-3.733	-3.762	-3.759	-3.777	-3.621	-3.658	-3.684	-3.677
4	-3.368	-3.421	-3.448	-3.443	-3.459	-3.310	-3.345	-3.370	-3.363
5	-3.119	-3.173	-3.199	-3.192	-3.206	-3.063	-3.097	-3.122	-3.115
6	-2.912	-2.967	-2.990	-2.982	-2.995	-2.858	-2.891	-2.914	-2.908
7	-2.734	-2.788	-2.811	-2.801	-2.813	-2.682	-2.713	-2.736	-2.729
8	-2.576	-2.630	-2.652	-2.641	-2.651	-2.526	-2.555	-2.578	-2.571
9	-2.434	-2.488	-2.508	-2.496	-2.506	-2.386	-2.414	-2.436	-2.430
10	-2.305	-2.358	-2.378	-2.364	-2.373	-2.259	-2.286	-2.306	-2.301
11	-2.185	-2.238	-2.257	-2.243	-2.251	-2.141	-2.167	-2.187	-2.182
12	-2.075	-2.126	-2.145	-2.129	-2.137	-2.033	-2.057	-2.077	-2.071
13	-1.971	-2.021	-2.039	-2.023	-2.030	-1.931	-1.954	-1.973	-1.968
14	-1.873	-1.922	-1.940	-1.923	-1.929	-1.835	-1.857	-1.875	-1.870
15	-1.781	-1.828	-1.846	-1.828	-1.834	-1.744	-1.765	-1.783	-1.778
16	-1.693	-1.738	-1.756	-1.737	-1.743	-1.657	-1.678	-1.694	-1.690
17	-1.608	-1.652	-1.670	-1.650	-1.656	-1.575	-1.594	-1.610	-1.606
18	-1.527	-1.570	-1.587	-1.567	-1.572	-1.496	-1.514	-1.529	-1.525
19	-1.450	-1.490	-1.507	-1.487	-1.492	-1.419	-1.437	-1.451	-1.448
20	-1.374	-1.413	-1.430	-1.409	-1.414	-1.346	-1.362	-1.376	-1.373
21	-1.302	-1.338	-1.356	-1.334	-1.338	-1.275	-1.290	-1.303	-1.300
22	-1.231	-1.266	-1.283	-1.261	-1.265	-1.206	-1.220	-1.233	-1.230
23	-1.162	-1.195	-1.213	-1.190	-1.194	-1.138	-1.152	-1.164	-1.161
24	-1.095	-1.126	-1.144	-1.121	-1.125	-1.073	-1.086	-1.097	-1.094
25	-1.030	-1.059	-1.077	-1.054	-1.057	-1.009	-1.021	-1.032	-1.029
26	- .966	- .993	-1.011	- .987	- .991	- .946	- .958	- .968	- .965
27	- .903	- .928	- .946	- .923	- .925	- .885	- .895	- .905	- .902
28	- .841	- .864	- .883	- .859	- .862	- .825	- .834	- .843	- .841
29	- .781	- .802	- .821	- .796	- .799	- .766	- .775	- .783	- .781
30	- .721	- .740	- .759	- .735	- .737	- .708	- .716	- .723	- .721
31	- .662	- .680	- .699	- .674	- .676	- .650	- .657	- .664	- .663
32	- .604	- .620	- .639	- .614	- .616	- .594	- .600	- .606	- .605
33	- .547	- .560	- .580	- .555	- .557	- .538	- .543	- .549	- .547
34	- .490	- .502	- .521	- .496	- .498	- .482	- .487	- .492	- .491
35	- .434	- .444	- .463	- .438	- .440	- .427	- .431	- .436	- .435
36	- .378	- .386	- .406	- .381	- .382	- .373	- .376	- .380	- .379
37	- .323	- .329	- .349	- .324	- .325	- .319	- .322	- .325	- .324
38	- .267	- .272	- .292	- .267	- .268	- .265	- .267	- .270	- .269
39	- .213	- .215	- .236	- .210	- .211	- .212	- .213	- .215	- .214
40	- .158	- .159	- .180	- .154	- .155	- .159	- .159	- .161	- .160
41	- .104	- .103	- .124	- .098	- .099	- .105	- .105	- .106	- .106
42	- .049	- .047	- .068	- .042	- .043	- .052	- .052	- .052	- .051
43	.005	.009	- .012	.013	.013	.001	.002	.002	.003
44	.059	.065	.044	.069	.069	.053	.056	.056	.057

APPENDIX A 3 Continued

SCORE GROUP	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
45	.113	.120	.100	.125	.125	.107	.109	.111	.111
46	.168	.176	.156	.181	.181	.160	.163	.165	.165
47	.222	.232	.212	.237	.237	.213	.217	.220	.220
48	.277	.289	.269	.293	.294	.266	.271	.274	.274
49	.332	.345	.326	.349	.350	.320	.326	.329	.329
50	.387	.402	.383	.406	.407	.374	.381	.385	.384
51	.443	.459	.440	.463	.465	.429	.436	.440	.440
52	.499	.517	.498	.520	.522	.484	.492	.497	.496
53	.556	.575	.557	.578	.580	.539	.548	.553	.553
54	.613	.633	.616	.637	.639	.595	.604	.611	.610
55	.671	.693	.676	.696	.699	.652	.662	.669	.668
56	.729	.753	.736	.756	.759	.710	.720	.727	.726
57	.789	.813	.798	.817	.820	.768	.779	.787	.786
58	.849	.875	.860	.878	.882	.827	.839	.848	.846
59	.911	.938	.924	.941	.944	.888	.900	.909	.908
60	.973	1.001	.989	1.004	1.008	.949	.962	.972	.970
61	1.037	1.066	1.055	1.069	1.074	1.012	1.025	1.036	1.034
62	1.102	1.133	1.122	1.135	1.140	1.076	1.090	1.101	1.099
63	1.169	1.200	1.191	1.203	1.208	1.141	1.156	1.168	1.166
64	1.237	1.270	1.262	1.272	1.278	1.209	1.224	1.237	1.234
65	1.307	1.341	1.335	1.344	1.349	1.278	1.294	1.307	1.304
66	1.379	1.414	1.411	1.417	1.423	1.349	1.366	1.380	1.377
67	1.454	1.490	1.488	1.493	1.499	1.423	1.441	1.455	1.452
68	1.531	1.568	1.569	1.571	1.577	1.499	1.518	1.532	1.529
69	1.611	1.650	1.653	1.652	1.659	1.578	1.597	1.613	1.609
70	1.695	1.734	1.740	1.736	1.744	1.661	1.681	1.697	1.693
71	1.782	1.822	1.832	1.824	1.832	1.747	1.768	1.784	1.780
72	1.874	1.915	1.928	1.917	1.925	1.837	1.859	1.876	1.872
73	1.970	2.012	2.029	2.014	2.023	1.933	1.955	1.974	1.969
74	2.073	2.116	2.137	2.117	2.126	2.035	2.058	2.077	2.071
75	2.182	2.226	2.253	2.227	2.237	2.143	2.167	2.187	2.181
76	2.299	2.344	2.378	2.345	2.355	2.259	2.284	2.305	2.299
77	2.427	2.473	2.514	2.473	2.484	2.386	2.411	2.433	2.426
78	2.566	2.613	2.663	2.614	2.625	2.525	2.551	2.574	2.567
79	2.722	2.769	2.830	2.769	2.781	2.679	2.707	2.730	2.722
80	2.898	2.946	3.020	2.946	2.958	2.855	2.883	2.907	2.899
81	3.102	3.151	3.241	3.151	3.163	3.058	3.087	3.112	3.103
82	3.347	3.397	3.508	3.396	3.409	3.302	3.332	3.358	3.349
83	3.656	3.707	3.846	3.705	3.719	3.610	3.641	3.669	3.659
84	4.083	4.135	4.311	4.133	4.147	4.037	4.069	4.097	4.086
85	4.798	4.850	5.079	4.848	4.863	4.751	4.784	4.813	4.802

APPENDIX A 4

QUANTITATIVE ABILITY - STANDARD ERRORS OF LOG ABILITY ESTIMATES, FIRST SPLIT

SCORE GROUP	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
1	1.044	1.048	1.048	1.039	1.041	1.040	1.049	1.051	1.051
2	.757	.763	.763	.752	.755	.753	.763	.765	.765
3	.633	.639	.639	.628	.631	.629	.638	.641	.640
4	.560	.568	.567	.556	.559	.556	.565	.568	.567
5	.512	.519	.519	.508	.511	.507	.516	.519	.518
6	.476	.485	.484	.473	.476	.472	.481	.483	.483
7	.450	.458	.457	.447	.450	.445	.453	.456	.455
8	.428	.437	.436	.426	.429	.424	.432	.434	.434
9	.411	.420	.419	.409	.412	.407	.414	.416	.416
10	.397	.405	.404	.395	.398	.393	.399	.401	.401
11	.384	.393	.392	.383	.386	.381	.387	.389	.389
12	.374	.383	.382	.374	.376	.370	.376	.378	.378
13	.365	.374	.373	.365	.368	.362	.367	.369	.369
14	.358	.366	.366	.358	.360	.354	.359	.361	.361
15	.351	.360	.359	.351	.354	.348	.352	.354	.354
16	.345	.354	.353	.346	.348	.342	.346	.348	.348
17	.340	.349	.348	.341	.343	.337	.341	.343	.343
18	.336	.344	.343	.337	.339	.333	.337	.339	.338
19	.332	.340	.340	.334	.335	.329	.333	.335	.335
20	.329	.337	.336	.331	.332	.326	.330	.332	.331
21	.327	.334	.334	.328	.330	.324	.327	.329	.328
22	.324	.332	.331	.326	.328	.321	.325	.326	.326
23	.323	.330	.329	.325	.326	.320	.323	.325	.324
24	.322	.329	.328	.324	.325	.319	.322	.323	.323
25	.321	.328	.327	.323	.324	.318	.321	.322	.322
26	.320	.327	.327	.323	.324	.318	.320	.322	.321
27	.321	.327	.327	.323	.324	.318	.320	.322	.321
28	.321	.328	.327	.324	.325	.318	.321	.322	.322
29	.322	.329	.328	.325	.326	.319	.322	.323	.323
30	.324	.330	.329	.326	.327	.321	.323	.325	.324
31	.326	.332	.332	.329	.330	.323	.326	.326	.326
32	.329	.335	.334	.331	.332	.325	.328	.329	.328
33	.332	.338	.337	.335	.336	.329	.331	.330	.332
34	.336	.342	.342	.339	.340	.332	.335	.336	.335
35	.341	.347	.346	.344	.345	.337	.340	.341	.340
36	.347	.353	.352	.350	.351	.343	.346	.346	.346
37	.353	.360	.359	.357	.358	.349	.352	.353	.352
38	.352	.368	.367	.365	.366	.357	.360	.361	.360
39	.371	.378	.377	.375	.376	.367	.370	.370	.370
40	.383	.389	.389	.387	.388	.378	.381	.381	.381
41	.397	.404	.403	.401	.402	.392	.395	.395	.394
42	.414	.421	.421	.419	.420	.409	.412	.412	.411

APPENDIX A 4 - Continued

SCORE GROUP	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
43	.435	.442	.442	.440	.441	.429	.433	.432	.432
44	.462	.469	.469	.468	.469	.456	.459	.459	.458
45	.497	.505	.505	.503	.505	.491	.494	.494	.493
46	.546	.555	.555	.553	.554	.540	.543	.543	.542
47	.619	.628	.629	.626	.628	.613	.616	.616	.615
48	.744	.753	.754	.752	.754	.738	.741	.741	.740
49	1.032	1.041	1.042	1.039	1.042	1.026	1.029	1.029	1.028

APPENDIX A 5

VERBAL ABILITY - STANDARD ERRORS OF LOG ABILITY ESTIMATES, FIRST SPLIT

SCORE GROUP	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
1	1.052	1.040	1.141	1.109	1.054	1.028	1.042	1.054	1.062
2	.761	.753	.825	.812	.765	.739	.752	.763	.771
3	.632	.628	.676	.674	.639	.612	.624	.633	.639
4	.554	.553	.588	.591	.563	.536	.547	.555	.560
5	.501	.503	.529	.534	.512	.485	.495	.501	.506
6	.462	.467	.486	.493	.475	.447	.456	.462	.466
7	.433	.438	.454	.462	.446	.418	.426	.432	.435
8	.409	.415	.429	.436	.423	.395	.403	.407	.411
9	.389	.397	.409	.416	.405	.376	.383	.388	.391
10	.373	.381	.392	.399	.389	.360	.367	.371	.374
11	.359	.368	.378	.385	.376	.346	.353	.357	.360
12	.347	.356	.365	.372	.364	.335	.341	.345	.348
13	.337	.346	.355	.361	.354	.325	.331	.335	.338
14	.328	.337	.345	.352	.346	.316	.322	.326	.328
15	.320	.329	.337	.343	.338	.308	.314	.318	.320
16	.313	.322	.330	.335	.331	.301	.307	.310	.313
17	.306	.316	.323	.329	.324	.295	.300	.304	.306
18	.301	.310	.317	.322	.318	.289	.295	.298	.300
19	.295	.305	.312	.317	.313	.284	.290	.293	.295
20	.291	.300	.307	.311	.308	.280	.285	.288	.290
21	.287	.296	.302	.306	.304	.276	.281	.284	.286
22	.283	.292	.298	.302	.300	.272	.277	.280	.282
23	.279	.288	.295	.298	.296	.269	.274	.277	.279
24	.276	.285	.291	.294	.292	.266	.271	.274	.275
25	.273	.282	.288	.291	.289	.263	.268	.271	.272
26	.271	.279	.285	.287	.286	.261	.265	.268	.270
27	.268	.276	.282	.285	.283	.259	.263	.266	.267
28	.266	.274	.279	.282	.281	.257	.261	.264	.265
29	.264	.272	.277	.279	.278	.255	.259	.262	.263
30	.262	.270	.275	.277	.276	.253	.257	.260	.262
31	.261	.268	.273	.275	.274	.252	.256	.259	.260
32	.260	.267	.271	.273	.273	.251	.255	.257	.259
33	.258	.266	.270	.271	.271	.250	.254	.256	.258
34	.257	.264	.268	.270	.270	.249	.253	.255	.257
35	.257	.263	.267	.269	.268	.249	.252	.255	.256
36	.256	.263	.266	.267	.267	.248	.252	.254	.255
37	.255	.262	.265	.266	.266	.248	.251	.253	.255
38	.255	.261	.264	.266	.266	.248	.251	.253	.254
39	.255	.261	.264	.265	.265	.248	.251	.253	.254
40	.255	.261	.264	.265	.265	.248	.251	.253	.254
41	.255	.261	.263	.264	.264	.248	.251	.253	.254
42	.255	.261	.263	.264	.264	.249	.252	.254	.255
43	.256	.262	.264	.264	.264	.249	.252	.254	.255

APPENDIX A 5 - Continued

SCORE GROUP	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
44	.256	.262	.264	.265	.265	.250	.253	.255	.256
45	.257	.263	.265	.265	.265	.251	.254	.256	.257
46	.258	.264	.265	.266	.266	.253	.255	.257	.258
47	.260	.266	.266	.267	.267	.254	.257	.258	.259
48	.261	.267	.268	.268	.268	.256	.258	.260	.261
49	.263	.269	.269	.269	.270	.257	.260	.262	.262
50	.265	.271	.271	.271	.271	.260	.262	.264	.264
51	.267	.273	.273	.273	.273	.262	.264	.266	.267
52	.270	.276	.275	.275	.276	.265	.267	.268	.269
53	.272	.279	.278	.278	.278	.268	.270	.271	.272
54	.276	.282	.281	.281	.281	.271	.273	.275	.275
55	.279	.286	.284	.284	.284	.275	.277	.278	.279
56	.283	.291	.288	.288	.288	.279	.281	.282	.283
57	.288	.295	.292	.292	.293	.283	.285	.287	.287
58	.293	.301	.297	.297	.297	.289	.290	.292	.292
59	.298	.307	.302	.302	.303	.294	.296	.297	.298
60	.305	.314	.309	.309	.309	.301	.303	.304	.304
61	.312	.318	.316	.316	.316	.308	.310	.311	.312
62	.320	.322	.324	.324	.324	.317	.318	.319	.320
63	.330	.332	.333	.333	.333	.326	.328	.329	.329
64	.341	.343	.344	.344	.344	.337	.339	.340	.340
65	.353	.355	.357	.357	.357	.350	.352	.353	.353
66	.368	.370	.372	.371	.372	.365	.367	.368	.368
67	.387	.389	.390	.390	.390	.384	.385	.386	.387
68	.409	.411	.412	.412	.412	.406	.408	.409	.409
69	.437	.439	.440	.440	.440	.435	.436	.437	.437
70	.474	.475	.476	.476	.477	.471	.473	.473	.474
71	.524	.526	.527	.526	.527	.522	.523	.524	.524
72	.599	.600	.601	.601	.601	.597	.598	.598	.599
73	.726	.727	.728	.728	.728	.724	.725	.725	.726
74	1.016	1.017	1.018	1.018	1.018	1.013	1.015	1.015	1.016

APPENDIX A 6

SCIENCE - STANDARD ERRORS OF LOG ABILITY ESTIMATES
FIRST SPLIT

SCORE GROUP	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
1	1.018	1.017	1.018	1.020	1.022	1.016	1.017	1.018	1.018
2	.728	.728	.729	.730	.732	.726	.728	.729	.729
3	.601	.601	.602	.604	.606	.599	.601	.602	.602
4	.526	.526	.527	.529	.531	.524	.526	.526	.526
5	.475	.476	.477	.478	.480	.473	.475	.476	.476
6	.438	.439	.440	.441	.443	.436	.438	.438	.438
7	.410	.411	.412	.413	.414	.407	.409	.410	.410
8	.387	.388	.389	.390	.392	.384	.386	.387	.387
9	.368	.370	.370	.372	.373	.365	.367	.368	.368
10	.352	.354	.355	.356	.357	.350	.351	.352	.352
11	.339	.341	.342	.343	.344	.336	.338	.339	.339
12	.327	.330	.330	.331	.332	.325	.326	.327	.327
13	.317	.320	.320	.321	.322	.315	.316	.317	.317
14	.308	.312	.311	.312	.313	.306	.307	.308	.308
15	.300	.304	.304	.305	.305	.298	.299	.301	.300
16	.293	.297	.297	.298	.298	.291	.292	.294	.293
17	.287	.291	.291	.291	.292	.284	.286	.287	.287
18	.282	.285	.285	.286	.286	.279	.280	.282	.281
19	.276	.280	.280	.281	.281	.274	.275	.276	.276
20	.272	.276	.275	.276	.277	.269	.271	.272	.271
21	.268	.272	.271	.272	.272	.265	.266	.268	.267
22	.264	.268	.267	.268	.268	.261	.263	.264	.263
23	.260	.265	.264	.264	.265	.258	.259	.260	.260
24	.257	.261	.261	.261	.262	.254	.256	.257	.257
25	.254	.259	.258	.258	.259	.251	.253	.254	.254
26	.252	.256	.255	.256	.256	.249	.250	.252	.251
27	.249	.254	.253	.253	.254	.246	.248	.249	.249
28	.247	.251	.251	.251	.251	.244	.246	.247	.247
29	.245	.249	.249	.249	.249	.242	.244	.245	.245
30	.243	.248	.247	.247	.247	.240	.242	.243	.243
31	.241	.246	.245	.245	.246	.239	.240	.241	.241
32	.240	.244	.244	.244	.244	.237	.239	.240	.240
33	.239	.243	.242	.242	.243	.236	.237	.239	.238
34	.237	.242	.241	.241	.241	.235	.236	.237	.237
35	.236	.241	.240	.240	.240	.234	.235	.236	.236
36	.235	.240	.239	.239	.239	.233	.234	.235	.235
37	.235	.239	.238	.238	.239	.232	.233	.235	.234
38	.234	.238	.238	.238	.238	.231	.233	.234	.234
39	.233	.238	.237	.237	.237	.231	.232	.233	.233
40	.233	.237	.237	.236	.237	.230	.232	.233	.233
41	.233	.237	.236	.236	.237	.230	.232	.233	.233
42	.233	.237	.236	.236	.236	.230	.231	.233	.232
43	.232	.236	.236	.236	.236	.230	.231	.232	.232

APPENDIX A 6 - Continued

SCORE GROUP	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
44	.233	.236	.236	.236	.236	.233	.231	.233	.232
45	.233	.236	.236	.236	.236	.230	.232	.233	.233
46	.233	.237	.237	.236	.237	.231	.232	.233	.233
47	.233	.237	.237	.236	.237	.231	.232	.233	.233
48	.234	.237	.238	.237	.237	.231	.233	.234	.234
49	.235	.238	.238	.237	.238	.232	.233	.235	.234
50	.235	.239	.239	.238	.239	.233	.234	.235	.235
51	.236	.240	.240	.239	.239	.234	.235	.236	.236
52	.237	.241	.241	.240	.240	.235	.236	.237	.237
53	.238	.242	.242	.241	.242	.236	.237	.239	.238
54	.240	.243	.244	.242	.243	.237	.239	.240	.240
55	.241	.244	.245	.244	.244	.239	.240	.241	.241
56	.243	.246	.247	.245	.246	.241	.242	.243	.243
57	.244	.247	.249	.246	.247	.242	.244	.245	.245
58	.246	.249	.251	.249	.249	.244	.246	.247	.246
59	.249	.251	.253	.251	.252	.247	.248	.249	.249
60	.251	.254	.256	.253	.254	.249	.250	.251	.251
61	.254	.256	.258	.256	.256	.252	.253	.254	.254
62	.256	.259	.261	.259	.259	.255	.256	.257	.257
63	.260	.262	.265	.262	.262	.258	.258	.260	.260
64	.263	.265	.268	.265	.266	.261	.262	.264	.263
65	.267	.269	.272	.269	.269	.265	.266	.267	.267
66	.271	.273	.276	.273	.273	.269	.270	.271	.271
67	.275	.278	.281	.277	.278	.274	.275	.276	.276
68	.280	.283	.287	.282	.283	.279	.280	.281	.281
69	.286	.288	.292	.288	.288	.284	.286	.287	.286
70	.292	.294	.299	.294	.294	.291	.292	.293	.292
71	.299	.301	.306	.300	.301	.297	.299	.300	.299
72	.307	.309	.314	.308	.309	.305	.306	.308	.307
73	.315	.317	.324	.317	.317	.314	.315	.316	.316
74	.325	.327	.334	.326	.327	.324	.325	.326	.326
75	.337	.338	.347	.338	.338	.335	.336	.338	.337
76	.350	.351	.361	.351	.352	.349	.350	.351	.350
77	.365	.367	.377	.366	.367	.364	.365	.366	.366
78	.384	.385	.397	.385	.385	.383	.384	.385	.384
79	.407	.408	.422	.407	.408	.406	.407	.408	.407
80	.435	.436	.452	.436	.436	.434	.435	.436	.436
81	.472	.473	.492	.473	.473	.471	.472	.473	.473
82	.523	.524	.546	.523	.524	.522	.523	.524	.523
83	.598	.599	.625	.598	.599	.597	.598	.599	.598
84	.725	.726	.755	.725	.726	.724	.725	.726	.725
85	1.014	1.015	1.047	1.015	1.015	1.014	1.015	1.015	1.015

APPENDIX A 7

QUANTITATIVE ABILITY - SCORING TABLES, SECOND SPLIT

SG	W	B	O	SG	W	B	O
1	-4.633	-4.492	-4.543	26	.146	.134	.134
2	-3.873	-3.743	-3.792	27	.250	.235	.236
3	-3.402	-3.282	-3.329	28	.354	.335	.339
4	-3.051	-2.940	-2.984	29	.458	.436	.442
5	-2.765	-2.663	-2.705	30	.564	.538	.546
6	-2.521	-2.427	-2.467	31	.671	.641	.651
7	-2.306	-2.220	-2.258	32	.779	.745	.758
8	-2.112	-2.033	-2.070	33	.889	.852	.866
9	-1.935	-1.863	-1.897	34	1.001	.961	.977
10	-1.770	-1.705	-1.738	35	1.116	1.072	1.091
11	-1.616	-1.557	-1.588	36	1.235	1.188	1.209
12	-1.471	-1.418	-1.446	37	1.359	1.307	1.331
13	-1.332	-1.285	-1.312	38	1.487	1.432	1.458
14	-1.200	-1.158	-1.183	39	1.622	1.562	1.591
15	-1.073	-1.036	-1.059	40	1.764	1.701	1.732
16	- .950	- .919	- .939	41	1.916	1.848	1.883
17	- .831	- .804	- .823	42	2.080	2.008	2.045
18	- .715	- .693	- .709	43	2.260	2.183	2.223
19	- .602	- .584	- .599	44	2.460	2.378	2.421
20	- .491	- .478	- .490	45	2.688	2.600	2.647
21	- .382	- .373	- .383	46	2.958	2.863	2.914
22	- .274	- .270	- .278	47	3.292	3.191	3.246
23	- .168	- .168	- .174	48	3.744	3.636	3.695
24	- .063	- .067	- .071	49	4.485	4.368	4.432
25	.042	.034	.032				

Key: SG: Score group
W : White
B : Black
O : Other

APPENDIX A 8

VERBAL ABILITY - SCORING TABLES, SECOND SPLIT

SG	W	B	O	SG	W	B	O
1	-5.176	-4.790	-4.720	38	.101	.061	.053
2	-4.370	-4.042	-3.991	39	.168	.123	.115
3	-3.876	-3.589	-3.552	40	.234	.184	.177
4	-3.516	-3.258	-3.231	41	.300	.246	.239
5	-3.230	-2.996	-2.977	42	.366	.308	.301
6	-2.992	-2.777	-2.763	43	.432	.371	.364
7	-2.787	-2.588	-2.579	44	.499	.434	.427
8	-2.606	-2.421	-2.416	45	.566	.497	.490
9	-2.443	-2.272	-2.269	46	.633	.561	.554
10	-2.294	-2.135	-2.134	47	.701	.625	.619
11	-2.157	-2.009	-2.010	48	.770	.690	.684
12	-2.029	-1.892	-1.895	49	.839	.756	.750
13	-1.909	-1.782	-1.786	50	.910	.823	.818
14	-1.796	-1.679	-1.684	51	.981	.891	.886
15	-1.688	-1.581	-1.586	52	1.054	.961	.956
16	-1.585	-1.487	-1.493	53	1.128	1.032	1.028
17	-1.487	-1.398	-1.404	54	1.203	1.104	1.101
18	-1.393	-1.311	-1.319	55	1.281	1.179	1.175
19	-1.301	-1.228	-1.236	56	1.360	1.256	1.252
20	-1.213	-1.148	-1.156	57	1.442	1.335	1.332
21	-1.128	-1.070	-1.078	58	1.527	1.417	1.414
22	-1.045	- .994	-1.003	59	1.615	1.502	1.500
23	- .964	- .920	- .929	60	1.706	1.590	1.589
24	- .885	- .848	- .857	61	1.801	1.683	1.682
25	- .807	- .777	- .786	62	1.901	1.780	1.780
26	- .732	- .708	- .717	63	2.007	1.883	1.883
27	- .657	- .640	- .649	64	2.119	1.993	1.994
28	- .584	- .573	- .582	65	2.239	2.111	2.112
29	- .512	- .507	- .516	66	2.369	2.239	2.240
30	- .441	- .442	- .451	67	2.511	2.379	2.381
31	- .371	- .377	- .386	68	2.669	2.534	2.537
32	- .302	- .313	- .322	69	2.847	2.710	2.713
33	- .234	- .250	- .259	70	3.053	2.913	2.917
34	- .166	- .187	- .196	71	3.299	3.157	3.162
35	- .098	- .125	- .134	72	3.610	3.466	3.472
36	- .032	- .063	- .071	73	4.038	3.892	3.899
37	.035	- .001	- .009	74	4.754	4.606	4.613

APPENDIX A 9

SCIENCE - SCORING TABLES, SECOND SPLIT

SG	W	B	O	SG	W	B	O
1	-4.853	-4.753	-4.783	44	.060	.061	.056
2	-4.131	-4.035	-4.067	45	.114	.114	.110
3	-3.698	-3.606	-3.639	46	.169	.166	.163
4	-3.383	-3.294	-3.329	47	.224	.219	.217
5	-3.134	-3.048	-3.084	48	.279	.271	.271
6	-2.926	-2.843	-2.879	49	.334	.324	.326
7	-2.746	-2.666	-2.703	50	.390	.378	.381
8	-2.588	-2.510	-2.547	51	.446	.431	.436
9	-2.445	-2.370	-2.407	52	.502	.486	.491
10	-2.316	-2.242	-2.280	53	.559	.540	.547
11	-2.196	-2.125	-2.162	54	.617	.595	.604
12	-2.085	-2.016	-2.053	55	.675	.651	.661
13	-1.981	-1.914	-1.951	56	.734	.708	.719
14	-1.882	-1.818	-1.854	57	.794	.765	.778
15	-1.789	-1.727	-1.763	58	.854	.823	.838
16	-1.701	-1.641	-1.676	59	.916	.882	.899
17	-1.616	-1.558	-1.592	60	.979	.943	.961
18	-1.535	-1.479	-1.513	61	1.043	1.004	1.024
19	-1.457	-1.403	-1.436	62	1.108	1.067	1.088
20	-1.381	-1.329	-1.361	63	1.175	1.131	1.154
21	-1.308	-1.258	-1.289	64	1.244	1.198	1.222
22	-1.237	-1.189	-1.220	65	1.314	1.265	1.292
23	-1.168	-1.122	-1.152	66	1.387	1.335	1.363
24	-1.100	-1.057	-1.085	67	1.462	1.407	1.437
25	-1.035	- .993	-1.021	68	1.539	1.482	1.514
26	- .970	- .931	- .957	69	1.620	1.560	1.593
27	- .907	- .870	- .895	70	1.703	1.641	1.676
28	- .845	- .810	- .834	71	1.791	1.726	1.763
29	- .784	- .751	- .774	72	1.883	1.815	1.854
30	- .724	- .693	- .715	73	1.980	1.909	1.950
31	- .665	- .636	- .657	74	2.082	2.008	2.051
32	- .607	- .580	- .600	75	2.192	2.115	2.160
33	- .549	- .524	- .543	76	2.310	2.230	2.277
34	- .492	- .469	- .487	77	2.438	2.354	2.403
35	- .435	- .415	- .431	78	2.578	2.491	2.543
36	- .379	- .361	- .376	79	2.733	2.644	2.697
37	- .324	- .307	- .321	80	2.910	2.817	2.873
38	- .268	- .254	- .267	81	3.114	3.018	3.076
39	- .213	- .201	- .213	82	3.359	3.260	3.320
40	- .158	- .148	- .159	83	3.669	3.567	3.629
41	- .104	- .096	- .105	84	4.096	3.991	4.055
42	- .049	- .043	- .051	85	4.811	4.702	4.770
43	.005	.009	.002				

APPENDIX A 10

QUANTITATIVE ABILITY - STANDARD ERRORS OF LOG ABILITY ESTIMATES, SECOND SPLIT

SG	W	B	O	SG	W	B	O
1	1.045	1.038	1.040	26	.321	.316	.319
2	.759	.751	.753	27	.321	.316	.319
3	.635	.627	.629	28	.322	.317	.320
4	.562	.554	.556	29	.323	.318	.321
5	.514	.506	.508	30	.325	.319	.322
6	.478	.470	.472	31	.327	.321	.325
7	.451	.443	.446	32	.329	.324	.327
8	.430	.422	.424	33	.333	.328	.331
9	.413	.405	.407	34	.337	.332	.335
10	.398	.390	.393	35	.341	.336	.340
11	.386	.378	.381	36	.347	.342	.345
12	.376	.368	.371	37	.354	.349	.352
13	.367	.359	.362	38	.362	.357	.360
14	.359	.352	.355	39	.372	.367	.370
15	.352	.345	.348	40	.383	.378	.382
16	.347	.340	.343	41	.397	.392	.396
17	.341	.335	.338	42	.414	.409	.412
18	.337	.331	.333	43	.436	.430	.434
19	.333	.327	.330	44	.463	.457	.460
20	.330	.324	.327	45	.498	.492	.496
21	.328	.321	.324	46	.547	.541	.545
22	.325	.319	.322	47	.620	.614	.618
23	.324	.318	.321	48	.745	.739	.743
24	.322	.317	.320	49	1.033	1.027	1.031
25	.322	.316	.319				

APPENDIX A 11

VERBAL ABILITY - STANDARD ERRORS OF LOG ABILITY ESTIMATES, SECOND SPLIT

SG	W	B	O	SG	W	B	O
1	1.071	1.035	1.023	38	.257	.248	.249
2	.778	.746	.734	39	.257	.248	.249
3	.645	.617	.608	40	.256	.249	.249
4	.565	.541	.533	41	.257	.249	.249
5	.510	.489	.482	42	.257	.249	.250
6	.470	.451	.445	43	.257	.250	.250
7	.439	.421	.416	44	.258	.251	.251
8	.414	.397	.394	45	.259	.252	.252
9	.394	.378	.375	46	.260	.253	.253
10	.378	.362	.359	47	.261	.254	.255
11	.364	.348	.346	48	.262	.256	.256
12	.352	.337	.335	49	.264	.258	.258
13	.341	.326	.325	50	.266	.260	.260
14	.332	.317	.316	51	.268	.262	.263
15	.324	.310	.308	52	.271	.265	.266
16	.316	.303	.301	53	.273	.268	.268
17	.310	.296	.295	54	.277	.271	.272
18	.304	.291	.290	55	.280	.275	.276
19	.299	.286	.285	56	.284	.279	.280
20	.294	.281	.281	57	.288	.284	.284
21	.290	.277	.277	58	.293	.289	.289
22	.286	.273	.273	59	.299	.295	.295
23	.282	.270	.270	60	.305	.301	.302
24	.279	.267	.267	61	.313	.309	.309
25	.276	.264	.264	62	.321	.317	.318
26	.273	.262	.262	63	.330	.327	.327
27	.271	.260	.259	64	.341	.338	.338
28	.269	.258	.258	65	.354	.351	.351
29	.267	.256	.256	66	.369	.366	.366
30	.265	.254	.254	67	.387	.384	.385
31	.263	.253	.253	68	.410	.407	.407
32	.262	.252	.252	69	.438	.435	.435
33	.260	.251	.251	70	.474	.472	.472
34	.259	.250	.250	71	.525	.522	.523
35	.258	.249	.249	72	.599	.597	.598
36	.258	.249	.249	73	.726	.724	.724
37	.257	.249	.249	74	1.016	1.014	1.014

APPENDIX A 12

SCIENCE - STANDARD ERRORS OF LOG ABILITY ESTIMATES, SECOND SPLIT

SG	W	B	O	SG	W	B	O
1	1.019	1.016	1.015	44	.233	.229	.231
2	.729	.726	.725	45	.233	.229	.232
3	.602	.599	.598	46	.234	.229	.232
4	.527	.524	.523	47	.234	.229	.232
5	.476	.473	.473	48	.234	.230	.233
6	.439	.436	.436	49	.235	.231	.233
7	.411	.407	.407	50	.236	.231	.234
8	.388	.384	.384	51	.237	.232	.235
9	.369	.366	.366	52	.238	.233	.236
10	.353	.350	.350	53	.239	.234	.237
11	.340	.336	.337	54	.240	.236	.239
12	.328	.325	.325	55	.242	.237	.240
13	.318	.315	.315	56	.243	.239	.242
14	.309	.306	.306	57	.245	.240	.243
15	.301	.298	.299	58	.247	.242	.245
16	.294	.291	.292	59	.249	.244	.247
17	.288	.284	.286	60	.252	.247	.250
18	.282	.279	.280	61	.254	.249	.252
19	.277	.274	.275	62	.257	.252	.255
20	.273	.269	.270	63	.260	.255	.258
21	.268	.265	.266	64	.264	.259	.262
22	.264	.261	.262	65	.267	.263	.266
23	.261	.257	.259	66	.271	.267	.270
24	.258	.254	.256	67	.276	.271	.274
25	.255	.251	.253	68	.281	.276	.279
26	.252	.248	.250	69	.286	.282	.285
27	.250	.246	.248	70	.293	.288	.291
28	.248	.244	.246	71	.299	.295	.298
29	.246	.242	.244	72	.307	.303	.306
30	.244	.240	.242	73	.316	.311	.314
31	.242	.238	.240	74	.326	.321	.324
32	.241	.236	.239	75	.337	.333	.336
33	.239	.235	.237	76	.350	.346	.349
34	.238	.234	.236	77	.366	.362	.364
35	.237	.233	.235	78	.384	.380	.383
36	.236	.232	.234	79	.407	.403	.406
37	.235	.231	.233	80	.435	.432	.434
38	.235	.230	.233	81	.472	.469	.471
39	.234	.230	.232	82	.523	.520	.522
40	.234	.229	.232	83	.598	.595	.597
41	.233	.229	.232	84	.725	.722	.724
42	.233	.229	.231	85	1.015	1.012	1.014
43	.233	.229	.231				

APPENDIX B

APPENDIX B 1

QUANTITATIVE ABILITY - LOG EASINESS ESTIMATES, FIRST SPLIT

ITEM	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
1	1.886	1.209	1.064	1.398	1.299	2.026	2.183	2.092	2.040
2	2.868	2.178	2.336	2.287	2.118	2.681	3.072	3.188	3.162
3	1.969	1.840	1.412	1.379	1.345	1.827	2.167	2.236	2.181
4	.596	.179	.308	.389	.410	.744	.727	.658	.578
5	2.281	1.995	2.892	2.459	2.535	2.039	2.270	2.317	2.325
6	1.558	2.178	1.838	1.943	2.221	1.269	1.453	1.568	1.505
7	2.069	2.686	2.490	1.959	2.221	1.875	2.010	2.061	2.117
8	1.300	1.586	1.472	1.448	1.455	1.331	1.339	1.239	1.292
9	1.646	1.995	1.930	1.600	1.579	1.578	1.610	1.664	1.672
10	1.367	2.401	2.267	2.010	2.194	1.142	1.091	1.272	1.370
11	1.040	1.061	1.355	1.351	1.234	1.013	.945	.990	1.002
12	.827	.395	.750	.401	.479	1.061	.931	.997	.859
13	.866	.930	1.569	1.157	1.133	.501	.668	.815	.813
14	.601	- .091	.100	.307	.149	.652	.771	.756	.733
15	.233	.114	.062	.036	.149	.128	.274	.246	.351
16	.631	.870	.642	.683	.587	.817	.645	.587	.597
17	- .614	- .431	- .462	- .524	- .621	- .829	- .648	- .629	- .754
18	.476	.657	.377	.490	.457	.488	.485	.503	.400
19	.285	.609	.702	.548	.447	.361	.230	.167	.076
20	- .094	- .409	- .366	- .266	- .153	- .044	- .098	- .053	.003
21	- .256	- .317	- .435	- .251	- .122	- .236	- .252	- .308	- .327
22	- .982	-1.259	-1.151	-1.330	-1.376	- .555	- .650	- .722	- .854
23	.983	.994	1.130	1.220	1.310	.861	.856	.954	.985
24	- .323	- .317	- .301	- .274	- .216	- .482	- .389	- .378	- .378
25	- .063	.994	.476	.451	.564	- .236	- .195	- .333	- .279
26	- .303	- .537	- .559	- .602	- .649	- .229	- .156	- .150	- .117
27	- .056	.146	- .046	- .016	- .054	.205	.029	- .102	- .223
28	.137	- .063	.149	.042	.024	.173	.099	.174	.224
29	- .913	- .636	-1.032	- .975	- .993	- .794	- .916	- .913	-1.031
30	- .816	-1.058	- .890	-1.030	-1.086	- .766	- .625	- .717	- .664
31	- .493	- .731	- .584	- .476	- .499	- .397	- .409	- .553	- .563
32	-1.273	-1.198	-1.304	-1.241	-1.316	-1.292	-1.349	-1.304	-1.390
33	- .212	.179	- .114	- .033	- .083	- .185	- .318	- .368	- .296
34	.407	1.061	1.022	.978	1.077	.083	.181	.234	.292
35	- .567	- .694	- .896	- .618	- .664	- .436	- .502	- .557	- .544
36	- .027	- .341	- .226	- .054	- .186	- .089	- .057	.051	.015
37	- .851	- .768	- .780	- .728	- .603	-1.035	-1.059	-1.060	- .955
38	- .739	- .768	- .603	- .622	- .757	- .916	- .773	- .785	- .826
39	-1.369	-1.318	-1.337	-1.304	-1.360	-1.387	-1.395	-1.507	-1.527
40	- .618	- .597	- .449	- .480	- .415	- .858	- .761	- .747	- .692
41	-1.150	-1.183	-1.378	-1.253	-1.260	- .945	-1.031	-1.206	-1.014
42	- .919	-1.074	- .970	- .878	- .986	- .916	- .988	- .949	- .953
43	-1.329	-1.289	-1.097	-1.056	-1.127	-1.614	-1.706	-1.690	-1.572
44	- .498	- .294	- .387	- .236	- .216	- .759	- .653	- .682	- .680
45	-1.002	- .960	-1.112	- .933	- .903	- .851	-1.163	-1.073	-1.156

APPENDIX B 1 - Continued

ITEM	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
46	-1.411	-1.847	-1.795	-1.733	-1.633	-1.005	-1.141	-1.209	-1.143
47	-1.217	-1.406	-1.392	-1.279	-1.234	-1.136	-1.298	-1.263	-1.221
48	-1.381	-1.434	-1.360	-1.239	-1.262	-1.414	-1.499	-1.631	-1.562
49	-2.103	-2.356	-2.370	-2.259	-2.228	-1.885	-2.057	-1.980	-1.984
50	-2.449	-2.883	-2.946	-2.845	-2.983	-1.564	-1.948	-1.900	-1.886

Key: The entries represent the log easiness estimates computed for each sub-group and for each item

OVERALL: log easiness estimates computed from the total sample of subjects

AMI: above median, income level I
:
BMIV: below median, income level IV

APPENDIX B 2

VERBAL ABILITY - LOG EASINESS ESTIMATES, FIRST SPLIT

ITEM	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
1	3.905	4.713	5.185	5.036	4.045	3.093	3.602	3.928	4.118
2	1.539	1.362	1.622	1.631	1.610	1.319	1.469	1.481	1.522
3	1.473	1.603	1.805	2.045	1.844	1.043	1.223	1.348	1.374
4	1.226	2.168	2.484	2.439	2.376	.755	.750	.958	.975
5	1.323	1.745	1.919	2.204	1.997	.848	.984	1.112	1.091
6	1.338	1.648	1.701	1.954	2.057	.835	1.020	1.135	1.204
7	1.368	1.648	1.577	1.737	1.853	1.440	1.216	1.173	1.158
8	.729	.570	.684	.618	.717	.775	.740	.700	.730
9	.880	1.101	1.117	1.117	1.162	.533	.731	.737	.705
10	- .283	- .455	- .457	- .486	- .478	- .155	- .145	- .262	- .260
11	.774	.631	.862	.806	.649	.917	.765	.750	.712
12	.349	.472	.790	.706	.810	.176	- .017	.098	.098
13	.106	.343	.245	.276	.440	- .017	- .199	- .176	- .084
14	- .260	.191	- .142	- .158	- .233	- .424	- .362	- .446	- .580
15	- .413	- .274	- .371	- .423	- .323	- .402	- .569	- .536	- .724
16	- .456	- .404	- .576	- .623	- .492	- .256	- .496	- .467	- .506
17	- .259	- .124	- .201	- .265	- .279	- .222	- .291	- .426	- .427
18	- .135	.143	.061	- .064	.069	- .346	- .375	- .345	- .427
19	- .560	- .517	- .687	- .583	- .616	- .555	- .640	- .595	- .706
20	- .668	- .650	- .751	- .869	- .698	- .525	- .562	- .816	- .682
21	- .721	- .757	- .683	- .894	- .825	- .548	- .622	- .798	- .770
22	-1.153	-1.301	-1.441	-1.411	-1.373	- .866	- .941	- .995	-1.105
23	- .611	- .429	- .544	- .699	- .588	- .623	- .612	- .765	- .803
24	- .541	- .391	- .754	- .749	- .774	- .416	- .397	- .392	- .552
25	- .741	- .793	- .739	- .756	- .651	- .841	- .933	- .921	- .998
26	-1.036	- .909	-1.122	-1.126	-1.101	- .954	-1.025	-1.112	-1.206
27	-1.144	-1.186	-1.294	-1.310	-1.156	- .972	-1.204	-1.286	-1.128
28	- .622	- .554	- .751	-1.004	- .885	- .284	- .490	- .507	- .431
29	-1.687	-1.633	-1.910	-1.849	-1.953	-1.234	-1.513	-1.591	-1.632
30	-1.304	-1.071	-1.456	-1.368	-1.355	-1.203	-1.378	-1.499	-1.616
31	1.977	2.416	2.190	2.415	2.259	1.542	1.838	1.898	2.011
32	2.300	3.123	2.711	3.439	2.773	1.551	2.054	2.239	2.415
33	1.796	1.517	1.735	1.613	1.771	1.534	1.813	1.885	1.894
34	.731	1.603	1.399	1.071	1.121	.533	.513	.510	.449
35	.324	1.225	1.290	1.024	1.107	- .102	- .226	- .188	- .046
36	1.389	3.030	2.669	2.744	2.696	.855	.921	1.081	1.218
37	1.393	1.291	1.577	1.419	1.531	.862	1.201	1.378	1.619
38	.810	.909	.806	.791	1.116	.741	.697	.693	.723
39	.081	.159	.007	.036	.019	.227	.190	.016	- .076
40	- .276	- .429	- .630	- .584	- .545	- .168	- .100	- .113	- .095
41	.031	.472	.295	.223	.350	- .142	- .232	- .252	- .264
42	- .051	- .366	- .363	- .410	- .475	.348	.211	.179	.076
43	- .205	- .207	- .367	- .233	- .284	- .162	- .226	- .251	- .268
44	- .476	- .578	- .792	- .763	- .841	.003	- .199	- .298	- .343
45	- .423	- .686	- .533	- .615	- .633	- .189	- .312	- .376	- .409

APPENDIX B 2 - Continued

ITEM	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
46	- .325	- .248	- .489	- .562	- .501	- .142	- .257	- .250	- .354
47	-1.210	-1.501	-1.430	-1.572	-1.571	- .824	- .842	- .918	- .922
48	- .648	- .851	-1.082	- .982	- .968	- .381	- .457	- .388	- .402
49	-1.097	- .863	-1.065	-1.160	-1.011	-1.192	-1.339	-1.373	-1.346
50	- .858	- .898	-1.025	-1.096	-1.092	- .367	- .622	- .829	- .687
51	-1.040	-1.036	-1.075	-1.175	-1.288	- .750	- .953	-1.017	-1.085
52	-1.232	- .990	-1.236	-1.308	-1.303	-1.203	-1.315	-1.414	-1.403
53	-1.083	-1.174	-1.383	-1.377	-1.347	- .990	- .854	- .790	- .927
54	-1.174	-1.105	-1.565	-1.604	-1.539	- .717	- .828	- .792	- .761
55	-1.541	-1.670	-1.706	-1.768	-1.881	-1.331	-1.277	-1.317	-1.356
56	1.802	1.907	1.880	2.214	2.176	1.014	1.485	1.843	1.966
57	1.454	1.696	1.653	1.784	1.771	.917	1.241	1.394	1.448
58	.204	- .353	- .355	- .382	- .389	.565	.507	.676	.641
59	.283	- .124	- .240	- .209	- .249	.643	.642	.591	.572
60	.650	.224	.169	.311	.339	.923	.937	.840	.726
61	.513	.674	.468	.592	.609	.393	.448	.424	.365
62	.151	- .138	.003	.044	.045	- .036	.138	.209	.279
63	.574	.884	.798	.770	.778	.208	.454	.449	.477
64	- .013	- .274	- .505	- .409	- .501	.373	.327	.262	.207
65	- .172	- .578	- .529	- .492	- .561	- .089	- .014	.097	.163
66	- .347	- .769	- .606	- .664	- .551	- .135	- .100	- .211	- .150
67	- .141	- .480	- .656	- .628	- .537	.284	.199	.153	.146
68	- .608	- .967	- .873	- .912	- .898	- .570	- .353	- .400	- .332
69	- .013	- .009	- .012	- .030	.002	- .452	- .136	- .048	- .018
70	- .935	-1.071	- .960	-1.094	-1.082	- .647	- .816	- .930	- .987
71	- .462	- .909	- .762	- .772	- .778	- .222	- .239	- .229	- .258
72	-1.557	-1.833	-1.962	-1.868	-1.785	-1.203	-1.164	-1.319	-1.307
73	-1.148	-1.128	-1.254	-1.247	-1.196	-1.203	-1.319	-1.194	-1.246
74	- .882	- .886	- .905	-1.070	-1.071	- .717	- .824	- .901	- .893
75	- .962	-1.278	-1.463	-1.434	-1.434	- .474	- .542	- .502	- .508

APPENDIX B 3

SCIENCE - LOG EASINESS ESTIMATES, FIRST SPLIT

ITEM	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
1	1.666	1.607	1.706	1.553	1.659	1.519	1.811	1.707	1.592
2	1.167	1.208	1.549	1.370	1.500	.926	1.049	1.097	1.014
3	1.607	1.459	1.673	1.581	1.629	1.538	1.651	1.649	1.556
4	.813	.499	.799	.690	.622	.815	.870	.941	.876
5	.167	- .158	.023	- .030	- .063	.243	.362	.247	.349
6	2.637	1.981	2.508	2.637	2.837	2.514	2.577	2.645	2.699
7	.151	- .096	- .017	- .128	- .111	.412	.314	.268	.306
8	1.059	1.065	1.074	.967	.912	1.101	1.134	1.051	1.171
9	.521	1.032	.865	.844	.928	.309	.454	.278	.318
10	.167	.252	.402	.419	.490	- .161	- .058	.023	.058
11	1.684	1.842	1.640	2.063	2.055	1.384	1.492	1.723	1.567
12	1.374	1.327	1.492	1.806	1.784	.968	1.156	1.354	1.330
13	- .229	- .409	- .284	- .397	- .531	.159	.040	- .051	- .218
14	- .656	-1.000	- .755	- .937	- .983	- .077	- .361	- .420	- .540
15	- .815	- .886	- .941	-1.135	-1.257	- .226	- .472	- .482	- .585
16	.824	.499	.543	.535	.603	.815	.883	1.012	1.067
17	.557	1.413	.951	1.099	.936	.351	.322	.310	.361
18	- .952	- .628	- .862	- .800	- .839	- .935	-1.165	-1.125	-1.220
19	1.317	1.607	1.706	1.570	1.395	1.274	1.191	1.233	1.240
20	1.320	1.171	1.300	1.374	1.384	1.131	1.246	1.314	1.393
21	1.299	2.058	1.927	1.840	1.925	.728	1.007	1.149	1.214
22	.667	.639	.698	.726	.756	.363	.569	.665	.690
23	.356	1.000	.577	.666	.784	.195	.051	.098	.239
24	.899	1.100	1.277	1.038	1.050	.669	.632	.837	.855
25	.528	1.100	.807	.658	.694	.339	.368	.391	.436
26	.905	1.459	1.336	1.417	1.176	.822	.666	.792	.726
27	.498	.371	.301	.233	.299	.624	.744	.562	.639
28	.235	.036	.130	.078	.225	.381	.277	.269	.217
29	.462	.214	.244	.370	.291	.586	.524	.549	.455
30	.350	.614	.591	.414	.364	.339	.354	.291	.231
31	.157	.455	.295	.388	.291	.112	.040	.009	.028
32	.232	- .047	.249	.172	.132	.297	.291	.317	.186
33	.148	- .235	.018	- .022	- .086	.375	.269	.286	.227
34	.556	.849	.856	.865	.932	.070	.266	.435	.578
35	- .079	- .158	- .149	- .343	- .336	.473	.247	.048	- .024
36	- .148	- .205	- .306	- .297	- .253	.047	.019	- .053	- .123
37	.098	- .047	.043	- .039	- .125	.106	.191	.208	.184
38	.212	.878	.477	.455	.386	- .018	.098	.110	.051
39	.186	.271	.183	.207	.296	.070	.010	.162	.172
40	.719	.714	.848	.912	1.064	.339	.515	.670	.721
41	- .053	- .279	- .031	- .068	- .029	- .184	- .066	- .037	- .100
42	- .120	- .047	- .056	- .187	- .107	.136	- .071	- .206	- .147
43	- .037	- .174	.028	- .168	- .235	.047	.109	.017	.101
44	- .151	- .174	.018	- .061	.099	- .390	- .341	- .289	- .201

APPENDIX B 3 Continued

ITEM	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
45	.033	- .174	- .012	.010	.014	.023	.029	.034	.051
46	- .298	- .294	- .310	- .433	- .330	- .101	- .234	- .292	- .233
47	- .135	- .250	- .149	- .209	- .084	- .137	- .153	- .166	- .083
48	- .400	- .395	- .441	- .510	- .393	- .202	- .355	- .427	- .398
49	- .312	- .437	- .523	- .488	- .529	- .274	- .204	- .150	- .159
50	- .151	- .031	- .036	- .022	.050	- .341	- .262	- .370	- .282
51	.020	.002	.068	- .034	- .208	.082	.183	.069	.110
52	- .439	- .588	- .547	- .516	- .511	- .371	- .414	- .398	- .472
53	- .245	- .437	- .292	- .195	- .245	- .341	- .333	- .219	- .303
54	- .618	- .746	- .634	- .511	- .524	- .782	- .766	- .716	- .682
55	- .601	- .912	- .949	- .936	- .922	- .208	- .234	- .335	- .320
56	1.012	.740	.873	.701	.612	1.274	1.298	1.243	1.072
57	-1.080	-1.151	-1.260	-1.343	-1.389	- .546	- .833	- .892	- .828
58	- .521	- .574	- .457	- .651	- .615	- .514	- .470	- .407	- .474
59	- .495	- .158	- .226	- .239	- .059	- .775	- .836	- .880	- .778
60	- .608	- .465	- .474	- .603	- .767	- .489	- .506	- .661	- .656
61	- .865	- .746	- .839	- .864	- .827	- .768	-1.035	- .969	- .982
62	- .524	- .265	- .292	- .335	- .357	- .629	- .650	- .760	- .764
63	- .520	- .493	- .437	- .592	- .520	- .701	- .546	- .572	- .472
64	-1.018	-1.125	-1.147	-1.219	-1.242	- .964	- .736	- .868	- .895
65	-1.066	- .720	-1.001	-1.087	-1.038	- .878	-1.161	-1.183	-1.232
66	- .948	-1.151	- .881	-1.028	- .981	- .935	- .937	- .944	- .876
67	-1.634	-1.493	-1.399	-1.568	-1.647	-1.592	-1.808	-1.901	-1.782
68	-1.381	-1.428	-1.449	-1.422	-1.468	-1.179	-1.388	-1.398	-1.335
69	-1.063	-1.213	-1.072	-1.068	-1.135	-1.123	-1.048	-1.054	-1.009
70	-1.592	-1.814	-1.753	-1.800	-1.696	-1.236	-1.432	-1.511	-1.398
71	-1.688	-1.637	-1.728	-1.639	-1.668	-1.741	-1.788	-1.813	-1.765
72	-1.914	-1.913	-1.874	-1.850	-1.964	-1.846	-1.909	-2.057	-1.951
73	-1.540	-1.545	-1.480	-1.503	-1.576	-1.582	-1.436	-1.630	-1.723
74	.690	1.065	.591	.726	.657	.599	.644	.732	.692
75	.734	1.246	1.025	.915	.965	.612	.572	.629	.632
76	- .109	.053	- .108	- .115	- .229	- .060	- .063	- .039	- .128
77	.488	.499	.433	.457	.436	.261	.442	.544	.625
78	.229	- .174	.094	.080	.079	.455	.365	.290	.359
79	.222	- .047	- .046	.065	.005	.142	.227	.426	.312
80	- .111	- .367	- .154	- .247	- .287	.177	- .011	- .024	- .062
81	-1.092	-1.188	- .949	-1.098	-1.140	-1.163	-1.102	-1.069	-1.116
82	.014	.053	.205	.178	.170	- .101	- .087	- .103	- .117
83	- .868	-1.050	- .888	- .973	-1.039	- .701	- .763	- .723	- .753
84	-1.342	-1.786	-1.511	-1.579	-1.616	-1.195	- .998	-1.009	-1.099
85	.212	- .080	.135	.145	.221	.035	.205	.258	.292
86	- .773	- .987	- .813	- .932	- .747	- .768	- .733	- .742	- .711

APPENDIX B 4

QUANTITATIVE ABILITY - STANDARD ERRORS OF LOG EASINESS ESTIMATES, FIRST SPLIT

ITEM	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
1	.030	.281	.146	.098	.106	.114	.087	.062	.076
2	.045	.442	.264	.149	.155	.143	.124	.097	.120
3	.031	.376	.171	.097	.108	.108	.086	.065	.080
4	.020	.184	.107	.064	.072	.086	.058	.041	.049
5	.035	.405	.345	.162	.189	.115	.090	.067	.084
6	.027	.442	.208	.126	.162	.094	.069	.052	.063
7	.032	.565	.284	.127	.162	.109	.082	.061	.078
8	.025	.334	.175	.100	.113	.095	.067	.047	.059
9	.028	.405	.217	.108	.120	.101	.072	.053	.066
10	.025	.492	.255	.130	.160	.092	.063	.047	.060
11	.023	.263	.166	.096	.103	.089	.061	.044	.054
12	.022	.199	.128	.064	.074	.090	.060	.044	.052
13	.022	.249	.183	.088	.098	.083	.058	.042	.052
14	.020	.167	.099	.062	.066	.085	.059	.042	.051
15	.019	.179	.098	.056	.066	.081	.055	.038	.047
16	.021	.242	.122	.072	.078	.087	.057	.040	.049
17	.017	.150	.083	.047	.051	.085	.054	.037	.045
18	.020	.222	.110	.066	.074	.083	.056	.040	.048
19	.019	.217	.125	.058	.073	.082	.054	.038	.046
20	.018	.151	.085	.051	.059	.081	.053	.037	.045
21	.018	.155	.083	.051	.060	.081	.053	.037	.045
22	.017	.124	.071	.040	.044	.083	.054	.037	.045
23	.023	.256	.150	.091	.106	.087	.060	.044	.054
24	.018	.155	.087	.051	.058	.082	.053	.037	.044
25	.018	.256	.114	.065	.077	.081	.053	.037	.045
26	.018	.145	.081	.046	.051	.081	.053	.037	.045
27	.018	.181	.094	.055	.061	.081	.054	.037	.045
28	.019	.168	.101	.056	.063	.081	.054	.038	.047
29	.017	.141	.072	.042	.047	.085	.055	.038	.046
30	.017	.129	.074	.042	.046	.084	.054	.037	.045
31	.017	.138	.080	.048	.053	.082	.053	.037	.045
32	.017	.125	.069	.040	.045	.093	.059	.040	.049
33	.018	.184	.092	.055	.060	.081	.053	.037	.045
34	.020	.263	.143	.081	.096	.081	.054	.038	.047
35	.017	.139	.074	.046	.051	.082	.053	.037	.044
36	.018	.154	.089	.054	.058	.081	.053	.038	.045
37	.017	.137	.076	.045	.052	.088	.056	.039	.046
38	.017	.137	.080	.046	.050	.086	.054	.037	.045
39	.017	.123	.069	.040	.044	.095	.060	.042	.050
40	.017	.143	.083	.048	.054	.086	.054	.037	.045
41	.017	.126	.069	.040	.045	.087	.056	.039	.046
42	.017	.128	.073	.043	.047	.086	.056	.038	.046
43	.017	.124	.072	.042	.046	.100	.064	.043	.050
44	.017	.156	.085	.051	.058	.084	.054	.037	.045

APPENDIX B 4 Continued

ITEM	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
45	.017	.131	.071	.043	.048	.085	.057	.039	.047
46	.017	.117	.066	.038	.043	.088	.057	.039	.047
47	.017	.122	.069	.040	.045	.090	.058	.040	.047
48	.017	.121	.069	.040	.045	.095	.061	.043	.050
49	.018	.118	.067	.038	.042	.108	.071	.047	.056
50	.019	.124	.072	.041	.045	.099	.069	.046	.054

APPENDIX B 5

VERBAL ABILITY - STANDARD ERRORS OF LOG EASINESS ESTIMATES, FIRST SPLIT

ITEM	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
1	.060	.498	.684	.395	.261	.158	.141	.117	.164
2	.023	.190	.124	.077	.082	.089	.064	.044	.057
3	.022	.210	.134	.093	.091	.085	.061	.043	.055
4	.021	.270	.183	.111	.116	.082	.056	.039	.050
5	.021	.224	.141	.100	.098	.083	.058	.041	.051
6	.021	.214	.128	.089	.100	.083	.058	.041	.053
7	.021	.214	.121	.081	.092	.091	.060	.041	.052
8	.018	.143	.086	.052	.058	.082	.056	.038	.048
9	.019	.172	.100	.062	.068	.081	.056	.038	.048
10	.016	.113	.064	.039	.042	.082	.055	.037	.047
11	.018	.146	.091	.055	.056	.083	.056	.038	.048
12	.017	.139	.089	.053	.060	.080	.055	.037	.046
13	.017	.134	.075	.046	.053	.081	.055	.037	.046
14	.016	.128	.068	.042	.044	.085	.056	.038	.049
15	.016	.116	.065	.040	.043	.085	.058	.039	.050
16	.016	.114	.063	.038	.042	.083	.057	.038	.048
17	.016	.119	.067	.041	.044	.083	.056	.038	.048
18	.016	.127	.071	.043	.047	.084	.056	.038	.048
19	.016	.112	.062	.039	.041	.087	.059	.039	.050
20	.016	.110	.062	.037	.041	.087	.058	.041	.049
21	.016	.109	.062	.037	.040	.087	.059	.040	.050
22	.017	.108	.060	.037	.039	.093	.063	.042	.054
23	.016	.113	.063	.038	.041	.088	.059	.040	.050
24	.016	.114	.062	.038	.040	.085	.057	.038	.048
25	.016	.109	.062	.038	.041	.092	.063	.041	.053
26	.016	.108	.060	.037	.039	.095	.064	.043	.055
27	.017	.108	.060	.037	.039	.095	.067	.045	.054
28	.016	.112	.062	.037	.040	.083	.057	.038	.048
29	.018	.111	.063	.038	.041	.103	.074	.050	.063
30	.017	.108	.061	.037	.039	.102	.071	.048	.063
31	.026	.302	.159	.110	.110	.094	.071	.050	.067
32	.030	.441	.203	.179	.140	.094	.076	.056	.077
33	.025	.203	.130	.076	.088	.093	.070	.050	.064
34	.018	.210	.113	.061	.067	.081	.055	.037	.047
35	.017	.180	.108	.060	.067	.082	.056	.037	.046
36	.022	.404	.199	.128	.135	.083	.057	.040	.053
37	.022	.185	.121	.070	.080	.083	.060	.043	.059
38	.019	.160	.089	.055	.067	.082	.056	.038	.048
39	.017	.127	.070	.044	.047	.080	.054	.037	.046
40	.016	.113	.062	.039	.042	.082	.055	.037	.046
41	.016	.139	.076	.046	.051	.082	.056	.037	.047
42	.016	.114	.065	.040	.042	.080	.054	.037	.046
43	.016	.117	.065	.041	.044	.082	.056	.037	.047
44	.016	.111	.061	.038	.040	.081	.055	.037	.047

APPENDIX B 5 Continued

ITEM	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
45	.016	.110	.063	.038	.041	.083	.056	.038	.048
46	.016	.117	.064	.039	.042	.082	.056	.037	.047
47	.017	.110	.060	.037	.039	.092	.061	.041	.052
48	.016	.109	.060	.037	.040	.085	.057	.038	.048
49	.017	.109	.060	.037	.040	.101	.070	.047	.058
50	.016	.109	.060	.037	.039	.084	.059	.041	.049
51	.016	.108	.060	.037	.039	.090	.063	.042	.054
52	.017	.108	.060	.037	.039	.102	.069	.047	.059
53	.017	.108	.060	.037	.039	.096	.061	.040	.052
54	.017	.108	.061	.037	.039	.090	.061	.040	.050
55	.018	.111	.062	.038	.040	.106	.068	.046	.058
56	.025	.240	.138	.100	.106	.085	.064	.049	.066
57	.022	.219	.125	.082	.088	.083	.061	.043	.056
58	.017	.115	.065	.040	.043	.081	.055	.038	.048
59	.017	.119	.066	.041	.044	.081	.055	.038	.047
60	.018	.129	.073	.047	.051	.083	.058	.039	.048
61	.018	.148	.080	.051	.056	.080	.055	.037	.047
62	.017	.119	.070	.044	.047	.081	.054	.037	.046
63	.018	.159	.089	.055	.059	.080	.055	.037	.047
64	.016	.116	.063	.040	.042	.080	.054	.037	.046
65	.016	.111	.063	.039	.041	.082	.055	.037	.046
66	.016	.109	.063	.038	.042	.082	.055	.037	.047
67	.016	.113	.062	.038	.042	.080	.054	.037	.046
68	.016	.108	.061	.037	.040	.087	.056	.038	.047
69	.016	.122	.070	.043	.046	.085	.055	.037	.046
70	.016	.108	.061	.037	.039	.089	.061	.041	.052
71	.016	.108	.062	.038	.040	.083	.056	.037	.047
72	.018	.114	.063	.038	.040	.102	.066	.046	.057
73	.017	.108	.060	.037	.039	.102	.069	.044	.056
74	.016	.109	.061	.037	.039	.090	.061	.041	.051
75	.016	.108	.061	.037	.039	.086	.058	.038	.048

APPENDIX B 6

SCIENCE - STANDARD ERRORS OF LOG EASINESS ESTIMATES, FIRST SPLIT

ITEM	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
1	.025	.230	.131	.073	.086	.097	.075	.052	.060
2	.021	.194	.122	.068	.080	.084	.060	.043	.051
3	.024	.215	.129	.074	.085	.097	.071	.051	.060
4	.019	.150	.091	.052	.057	.083	.058	.041	.049
5	.017	.125	.072	.042	.046	.078	.054	.037	.045
6	.037	.271	.189	.120	.148	.136	.101	.075	.093
7	.017	.127	.071	.041	.046	.079	.053	.037	.045
8	.020	.184	.101	.058	.063	.087	.061	.043	.053
9	.018	.181	.093	.055	.063	.078	.054	.037	.045
10	.017	.139	.079	.048	.054	.078	.052	.037	.044
11	.025	.254	.127	.092	.103	.093	.068	.052	.060
12	.022	.204	.119	.082	.091	.085	.062	.046	.056
13	.016	.120	.067	.039	.042	.078	.053	.036	.044
14	.016	.113	.063	.036	.040	.077	.053	.037	.044
15	.016	.113	.063	.036	.040	.078	.053	.037	.045
16	.019	.150	.083	.049	.056	.083	.058	.042	.052
17	.018	.211	.096	.061	.064	.078	.053	.037	.045
18	.016	.116	.062	.037	.041	.085	.060	.041	.049
19	.022	.230	.131	.074	.077	.091	.062	.045	.054
20	.022	.191	.110	.068	.076	.088	.063	.046	.057
21	.022	.280	.144	.083	.097	.082	.060	.044	.054
22	.018	.157	.088	.053	.059	.078	.055	.039	.048
23	.017	.179	.084	.052	.060	.078	.053	.037	.045
24	.019	.186	.109	.059	.067	.081	.055	.041	.049
25	.018	.186	.091	.051	.058	.078	.054	.038	.046
26	.019	.215	.112	.069	.070	.083	.056	.040	.048
27	.018	.144	.077	.045	.051	.080	.056	.039	.047
28	.017	.131	.073	.043	.050	.078	.053	.037	.045
29	.018	.137	.076	.047	.051	.080	.055	.038	.046
30	.017	.156	.084	.047	.052	.078	.054	.037	.045
31	.017	.148	.077	.047	.051	.077	.053	.037	.044
32	.017	.128	.076	.044	.049	.078	.053	.037	.044
33	.017	.123	.071	.042	.046	.078	.053	.037	.045
34	.018	.169	.093	.055	.064	.077	.053	.038	.047
35	.016	.125	.069	.039	.044	.079	.053	.037	.044
36	.016	.124	.067	.039	.044	.077	.053	.036	.044
37	.016	.128	.072	.042	.046	.077	.053	.037	.044
38	.017	.171	.081	.048	.052	.077	.053	.037	.044
39	.017	.140	.075	.045	.051	.077	.053	.037	.044
40	.019	.161	.092	.056	.067	.078	.055	.039	.048
41	.016	.122	.071	.041	.047	.078	.052	.037	.044
42	.016	.128	.070	.040	.046	.078	.052	.036	.044
43	.016	.125	.072	.041	.044	.077	.053	.037	.044

APPENDIX B 6 Continued

ITEM	OVERALL	AMI	AMII	AMIII	AMIV	BMI	BMII	BMIII	BMIV
44	.016	.125	.071	.042	.048	.079	.053	.037	.044
45	.016	.125	.071	.042	.047	.077	.053	.037	.044
46	.016	.122	.067	.039	.044	.078	.053	.037	.044
47	.016	.123	.069	.040	.046	.078	.052	.036	.044
48	.016	.120	.065	.038	.043	.078	.053	.037	.044
49	.016	.119	.064	.038	.042	.078	.053	.036	.044
50	.016	.129	.070	.042	.048	.078	.053	.037	.044
51	.016	.130	.072	.042	.045	.077	.053	.037	.044
52	.016	.117	.064	.038	.042	.079	.053	.037	.044
53	.016	.119	.067	.040	.044	.078	.053	.036	.044
54	.016	.115	.063	.038	.042	.083	.055	.038	.045
55	.016	.113	.062	.036	.040	.078	.053	.037	.044
56	.020	.163	.093	.052	.057	.091	.064	.045	.052
57	.016	.113	.062	.036	.040	.080	.056	.039	.046
58	.016	.117	.065	.037	.042	.080	.053	.037	.044
59	.016	.125	.068	.040	.046	.083	.056	.039	.046
60	.016	.119	.065	.038	.041	.080	.054	.038	.045
61	.016	.115	.062	.037	.041	.083	.058	.039	.047
62	.016	.123	.067	.039	.043	.081	.054	.038	.046
63	.016	.118	.065	.038	.042	.082	.054	.037	.044
64	.016	.113	.062	.036	.040	.086	.055	.039	.046
65	.016	.115	.062	.036	.040	.084	.060	.041	.050
66	.016	.113	.062	.036	.040	.085	.057	.039	.046
67	.017	.114	.062	.037	.040	.101	.071	.050	.058
68	.016	.114	.062	.036	.040	.090	.063	.043	.051
69	.016	.113	.062	.036	.040	.089	.058	.040	.047
70	.017	.119	.064	.038	.041	.091	.064	.045	.052
71	.017	.116	.064	.037	.041	.106	.071	.049	.057
72	.018	.120	.065	.038	.042	.110	.074	.053	.061
73	.017	.115	.062	.036	.040	.101	.064	.046	.057
74	.018	.184	.084	.053	.057	.080	.056	.040	.048
75	.019	.197	.099	.056	.064	.080	.055	.039	.047
76	.016	.132	.069	.041	.044	.077	.052	.037	.044
77	.018	.150	.080	.048	.053	.078	.054	.038	.047
78	.017	.125	.073	.043	.048	.079	.054	.037	.045
79	.017	.128	.070	.043	.047	.078	.053	.038	.045
80	.016	.120	.069	.040	.044	.078	.052	.037	.044
81	.016	.113	.062	.036	.040	.090	.059	.040	.048
82	.016	.132	.075	.044	.049	.078	.052	.036	.044
83	.016	.113	.062	.036	.040	.082	.055	.038	.045
84	.016	.118	.062	.037	.040	.090	.057	.040	.048
85	.017	.128	.074	.044	.050	.077	.053	.037	.045
86	.016	.113	.062	.036	.041	.083	.055	.038	.045

APPENDIX B 7

QUANTITATIVE ABILITY - LOG EASINESS ESTIMATES, SECOND SPLIT

ITEM	W	B	O	ITEM	W	B	O
1	1.851	2.067	1.832	26	- .317	- .005	- .236
2	2.934	2.608	2.752	27	- .117	.189	.153
3	2.009	1.726	1.866	28	.184	.055	- .057
4	.575	.766	.582	29	- .930	- .614	- .801
5	2.356	1.958	1.998	30	- .814	- .464	- .859
6	1.564	1.167	1.577	31	- .492	- .368	- .457
7	2.133	1.811	1.949	32	-1.295	-1.577	-1.174
8	1.284	1.359	1.408	33	- .236	- .165	- .109
9	1.713	1.526	1.375	34	.420	.107	.384
10	1.410	.946	1.358	35	- .589	- .320	- .609
11	1.005	.859	1.233	36	- .032	.055	- .160
12	.837	.986	.780	37	- .852	- .973	- .836
13	.910	.411	.699	38	- .760	- .656	- .711
14	.611	.986	.560	39	-1.384	-1.132	-1.332
15	.253	- .013	.080	40	- .641	- .833	- .521
16	.612	.891	.768	41	-1.143	-1.122	-1.135
17	- .580	-1.291	- .719	42	- .967	- .761	- .828
18	.478	.553	.486	43	-1.325	-1.465	-1.372
19	.286	.055	.404	44	- .492	- .614	- .437
20	- .087	- .513	- .173	45	-1.034	- .708	- .828
21	- .269	- .142	- .194	46	-1.430	- .954	-1.348
22	- .977	- .588	-1.155	47	-1.241	-1.122	-1.163
23	1.001	.766	.871	48	-1.391	-1.226	-1.364
24	- .337	- .273	- .236	49	-2.107	-2.136	-2.088
25	- .061	- .384	.071	50	-2.527	-1.429	-2.285

Key: W: White

B: Black

O: Others

APPENDIX B 8

VERBAL ABILITY - LOG EASINESS ESTIMATES, SECOND SPLIT

ITEM	W	B	O	ITEM	W	B	O
1	4.324	3.373	2.781	39	.089	.148	.182
2	1.610	1.343	1.059	40	- .267	- .298	- .300
3	1.591	1.078	.867	41	.018	.105	.032
4	1.323	.503	.889	42	- .087	.084	.141
5	1.438	.991	.794	43	- .229	- .126	- .127
6	1.423	.937	.912	44	- .536	.077	- .188
7	1.322	1.441	1.535	45	- .475	- .313	- .188
8	.700	.655	.961	46	- .365	- .134	- .249
9	.872	.728	.743	47	-1.247	- .819	-1.069
10	- .328	- .260	- .285	48	- .726	- .484	- .393
11	.775	.757	1.002	49	-1.144	-1.092	- .891
12	.401	- .003	.119	50	- .905	- .468	- .546
13	.085	.012	.006	51	-1.074	- .956	- .822
14	- .287	- .200	- .372	52	-1.286	-1.302	- .833
15	- .435	- .540	- .419	53	-1.132	- .854	- .895
16	- .488	- .351	- .390	54	-1.248	- .580	- .845
17	- .327	- .215	.017	55	-1.591	-1.112	-1.569
18	- .189	- .359	.130	56	1.999	.922	1.044
19	- .571	- .405	- .761	57	1.540	1.224	.952
20	- .745	- .290	- .422	58	.192	.496	.267
21	- .782	- .492	- .303	59	.231	.691	.504
22	-1.197	- .706	-1.009	60	.620	1.167	.885
23	- .639	- .572	- .484	61	.529	.461	.407
24	- .614	- .163	- .400	62	.158	.026	.094
25	- .790	- .984	- .517	63	.580	.205	.583
26	-1.083	- .810	- .708	64	- .016	.333	.086
27	-1.190	-1.013	- .946	65	- .144	- .252	- .411
28	- .695	- .405	- .357	66	- .377	- .163	- .138
29	-1.710	-1.527	-1.489	67	- .194	.354	.119
30	-1.351	-1.163	-1.160	68	- .565	- .828	- .950
31	2.069	1.552	1.780	69	.073	-1.072	- .282
32	2.521	1.610	1.679	70	- .976	- .540	- .772
33	1.876	1.543	1.481	71	- .484	- .397	- .343
34	.717	.439	.710	72	-1.578	-1.269	-1.550
35	.334	- .024	.286	73	-1.158	- .994	-1.077
36	1.432	.817	1.195	74	- .925	- .689	- .641
37	1.509	.953	.907	75	-1.005	- .548	- .780
38	.803	.742	.730				

APPENDIX B 9

SCIENCE - LOG EASINESS ESTIMATES, SECOND SPLIT

ITEM	W	B	O	ITEM	W	B	O
1	1.660	1.803	1.563	44	- .150	- .411	- .101
2	1.194	.898	1.148	45	- .016	.254	.181
3	1.609	1.641	1.549	46	- .326	- .083	- .206
4	.845	.787	.721	47	- .128	- .338	- .157
5	.198	.172	- .069	48	- .409	- .516	- .352
6	2.726	2.419	2.222	49	- .292	- .411	- .362
7	.096	.550	.381	50	- .161	- .259	- .147
8	1.066	.868	1.143	51	.011	.418	.085
9	.522	.117	.663	52	- .459	- .396	- .451
10	.201	- .309	.085	53	- .269	- .237	- .154
11	1.721	1.250	1.652	54	- .609	-1.095	- .706
12	1.446	.982	1.205	55	- .653	.000	- .516
13	- .221	.089	- .168	56	.987	1.372	1.077
14	- .695	- .118	- .537	57	-1.122	- .470	- .958
15	- .877	- .146	- .647	58	- .539	- .404	- .382
16	.850	.621	.685	59	- .516	- .679	- .461
17	.564	.418	.573	60	- .646	- .309	- .561
18	- .983	- .793	- .980	61	- .868	- .930	- .848
19	1.346	1.241	1.324	62	- .510	- .624	- .434
20	1.333	1.149	1.439	63	- .506	- .531	- .623
21	1.396	.543	.968	64	-1.065	- .878	- .915
22	.689	.322	.514	65	-1.120	- .818	- .834
23	.357	.021	.276	66	- .969	- .663	- .848
24	.882	.586	.860	67	-1.645	-1.370	-1.596
25	.513	.233	.703	68	-1.412	-1.153	-1.264
26	.892	.809	.983	69	-1.086	- .993	-1.041
27	.461	.621	.494	70	-1.650	-1.192	-1.462
28	.246	.041	.111	71	-1.716	-1.472	-1.639
29	.431	.743	.650	72	-1.927	-1.711	-1.898
30	.316	.301	.445	73	-1.566	- .984	-1.503
31	.155	- .041	.299	74	.725	.564	.556
32	.236	.158	.215	75	.755	.453	.637
33	.140	.192	.280	76	- .091	- .021	- .150
34	.598	- .104	.433	77	.529	- .055	.405
35	- .107	.579	- .005	78	.262	.543	.045
36	- .163	.083	- .101	79	.251	.131	.056
37	.088	.288	.133	80	- .134	.192	- .150
38	.223	- .083	.405	81	-1.092	-1.153	-1.001
39	.213	- .083	.118	82	.043	- .302	- .126
40	.755	- .281	.552	83	- .883	- .546	- .891
41	- .031	- .195	- .241	84	-1.329	- .912	-1.361
42	- .128	.110	- .026	85	.299	- .160	.031
43	- .014	- .181	- .083	86	- .749	- .719	- .912

APPENDIX B 10

QUANTITATIVE ABILITY - STANDARD ERRORS OF LOG EASINESS ESTIMATES
SECOND SPLIT

ITEM	W	B	O	ITEM	W	B	O
1	.040	.111	.105	26	.023	.088	.066
2	.063	.131	.148	27	.024	.088	.069
3	.043	.103	.106	28	.025	.088	.067
4	.027	.089	.075	29	.022	.093	.064
5	.049	.108	.111	30	.022	.092	.064
6	.036	.093	.096	31	.023	.091	.065
7	.045	.105	.109	32	.022	.113	.064
8	.033	.096	.092	33	.023	.089	.067
9	.038	.098	.091	34	.026	.088	.072
10	.035	.091	.090	35	.022	.090	.064
11	.030	.090	.087	36	.024	.088	.066
12	.029	.091	.078	37	.022	.099	.064
13	.030	.088	.076	38	.022	.094	.064
14	.027	.091	.074	39	.022	.102	.064
15	.025	.088	.068	40	.022	.097	.064
16	.027	.090	.078	41	.022	.102	.064
17	.022	.106	.064	42	.022	.095	.064
18	.027	.088	.073	43	.022	.110	.064
19	.025	.088	.072	44	.023	.093	.065
20	.024	.092	.066	45	.022	.095	.064
21	.023	.089	.066	46	.022	.099	.064
22	.022	.093	.064	47	.022	.102	.064
23	.030	.089	.079	48	.022	.104	.064
24	.023	.090	.066	49	.023	.134	.069
25	.024	.091	.068	50	.025	.109	.071

APPENDIX B 11

VERBAL ABILITY - STANDARD ERRORS OF LOG EASINESS ESTIMATES
SECOND SPLIT

ITEM	W	B	O	ITEM	W	B	O
1	.100	.180	.120	38	.024	.086	.065
2	.031	.094	.069	39	.022	.085	.061
3	.031	.090	.067	40	.021	.088	.061
4	.028	.085	.067	41	.021	.085	.061
5	.029	.089	.066	42	.021	.085	.061
6	.029	.088	.067	43	.021	.086	.061
7	.028	.095	.078	44	.021	.085	.061
8	.024	.086	.068	45	.021	.088	.061
9	.025	.086	.065	46	.021	.087	.061
10	.021	.087	.061	47	.022	.095	.064
11	.024	.087	.068	48	.021	.090	.061
12	.023	.086	.061	49	.021	.100	.063
13	.022	.086	.061	50	.021	.090	.061
14	.021	.087	.061	51	.021	.098	.062
15	.021	.091	.061	52	.022	.106	.062
16	.021	.088	.061	53	.021	.096	.063
17	.021	.087	.061	54	.022	.091	.062
18	.021	.088	.061	55	.023	.101	.070
19	.021	.089	.062	56	.035	.088	.069
20	.021	.088	.061	57	.030	.092	.068
21	.021	.090	.061	58	.022	.085	.062
22	.022	.093	.064	59	.022	.086	.063
23	.021	.091	.061	60	.024	.091	.067
24	.021	.087	.061	61	.023	.085	.063
25	.021	.098	.061	62	.022	.086	.061
26	.021	.095	.062	63	.023	.085	.064
27	.022	.099	.063	64	.021	.085	.061
28	.021	.089	.061	65	.021	.087	.061
29	.023	.112	.069	66	.021	.087	.061
30	.022	.102	.065	67	.021	.085	.061
31	.036	.098	.083	68	.021	.095	.063
32	.044	.099	.081	69	.022	.100	.061
33	.034	.097	.076	70	.021	.091	.062
34	.024	.085	.065	71	.021	.089	.061
35	.022	.086	.062	72	.023	.105	.069
36	.029	.087	.071	73	.021	.098	.064
37	.030	.088	.067	74	.021	.093	.061
				75	.021	.091	.062

APPENDIX B 12

SCIENCE - STANDARD ERRORS OF LOG EASINESS ESTIMATES, SECOND SPLIT

ITEM	W	B	O	ITEM	W	B	O
1	.033	.106	.085	44	.021	.086	.060
2	.028	.087	.075	45	.021	.083	.062
3	.032	.102	.084	46	.021	.084	.059
4	.025	.086	.067	47	.021	.086	.060
5	.022	.083	.060	48	.020	.088	.059
6	.051	.130	.107	49	.021	.086	.059
7	.022	.084	.063	50	.021	.085	.060
8	.027	.087	.075	51	.021	.084	.061
9	.023	.083	.067	52	.020	.086	.059
10	.022	.085	.061	53	.021	.085	.060
11	.034	.093	.087	54	.020	.098	.059
12	.031	.088	.076	55	.020	.084	.059
13	.021	.083	.060	56	.026	.095	.073
14	.020	.084	.059	57	.021	.087	.060
15	.020	.084	.059	58	.020	.086	.059
16	.025	.085	.067	59	.020	.090	.059
17	.024	.084	.065	60	.020	.085	.059
18	.020	.092	.060	61	.020	.094	.060
19	.030	.093	.079	62	.020	.089	.059
20	.029	.091	.081	63	.020	.088	.059
21	.030	.084	.071	64	.021	.093	.060
22	.024	.083	.065	65	.021	.092	.060
23	.023	.084	.062	66	.020	.090	.060
24	.026	.084	.070	67	.022	.105	.066
25	.023	.083	.067	68	.021	.099	.062
26	.026	.086	.072	69	.021	.096	.061
27	.023	.085	.064	70	.022	.100	.064
28	.022	.083	.061	71	.022	.108	.066
29	.023	.086	.066	72	.023	.116	.070
30	.022	.083	.064	73	.022	.095	.065
31	.022	.084	.063	74	.025	.084	.065
32	.022	.083	.062	75	.025	.084	.066
33	.022	.083	.062	76	.021	.084	.060
34	.024	.084	.064	77	.023	.084	.064
35	.021	.084	.060	78	.022	.084	.061
36	.021	.083	.060	79	.022	.083	.061
37	.022	.083	.061	80	.021	.083	.060
38	.022	.084	.064	81	.021	.099	.060
39	.022	.084	.061	82	.021	.085	.060
40	.025	.083	.065	83	.020	.088	.060
41	.021	.085	.059	84	.021	.094	.063
42	.021	.083	.060	85	.022	.084	.061
43	.021	.084	.060	86	.020	.090	.060

LIST OF REFERENCES

LIST OF REFERENCES

Andersen, E.B., Asymptotic properties of conditional maximum likelihood estimators, The Journal of the Royal Statistical Society, 1970, 32, Pp. 283-301.

Andersen, E.B., The numerical solution of a set of conditional estimation equations, The Journal of the Royal Statistical Society: Series B, 1972, 34(1), Pp. 42-54.

Andersen, E.B., A goodness of fit test for the Rasch model, Psychometrika, Vol. 38, No 1, 1973, Pp. 123-140.

Anderson, T.W., Some scaling models and estimation procedures in the latent class model, In U. Grenander (Ed.), Probability and statistics, New York: Wiley, 1959, Pp. 9-38.

Baker, F.B., Empirical comparison of item parameters based on the logistic and normal functions, Psychometrika, XXVI, 1961, 239.

Berkson, J., A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function, Journal of the American Statistical Association, XLIV, 1953, 565.

Birnbaum, A., Some latent trait models and their use in inferring an examinee's ability, Part V of F.M. Lord and M.R. Novick, Statistical Theories of Mental Test Scores, Reading, Mass: Addison-Wesley, 1968.

Bliss, C.I., The comparison of the dosage mortality curve, Annals of Applied Biology, XXII, 1935, 134.

Bock, R.D., and Jones, L.V., The Measurement and Prediction of Judgment and Choice, San Francisco: Holden Day, 1968.

Bramble, W.J., A least square method of parameter estimation for the logistic measurement model, 1969, (Unpublished).

Brink, N.E., Rasch's logistic model vs. the Guttman model, Educational and Psychological Measurement, 1972, 32, Pp. 921-927.

Cartledge, C.M., A comparison of equipercentile and Rasch equating methodologies, Doctoral dissertation, University of Georgia, 1974, (Unpublished).

Coffman, W.E., A factor analysis of the verbal sections of the Scholastic Aptitude Test, Research Bulletin 66-30, Princeton, N.J., Educational Testing Service, 1966.

Cramer, E.M., A comparison of three methods of fitting the normal ogive, Psychometrika, XXVII, 1962, 183.

Douglas, G.A., Test Design strategies for the Rasch psychometric model, Doctoral dissertation, University of Chicago, 1975, (Unpublished).

Erdmann, J.B., et al, The Medical College Admission Test: Past, Present, Future, Journal of Medical Education, Vol. 46, No. 11, 1971.

Finney, D.J., Probit Analysis, London: Cambridge University Press, 1952.

Gaddum, J.G., Reports on biological standards III, methods of biological assay depending on quantal response, Medical Research Council, Special Report Series No. 183, 1933.

Garwood, F., The application of maximum likelihood to dosage mortality curves, Biometrika, XXXII, 1941, 46.

Grey, D.R., and Morgan, B.J., Some aspects of ROC curve-fitting: normal and logistic models, J. of Math. Psych., 1972, 9, Pp.128-139.

Gulliksen, H., Theory of Mental Tests, New York: John Wiley and Sons, 1950.

Keesling, J.W., Computer programming of the model, Presentation at the 1969 AERA Presession on Person-Free Item Calibration and Item-Free Person Measurement, Los Angeles, California, (a), (Unpublished).

Keesling, J.W., Evaluation of fit control of the model, Presentation at the 1969 AERA Presession on Person-Free Item Calibration and Item-Free Person Measurement, Los Angeles, California, (b), (Unpublished).

Lord, F.M., A theory of test scores, Psychometric Monographs, No. 7, 1952.

Lord, F.M., An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability, Psychometrika, XVII, 1953, 57.

Lord, F.M., A strong true-score theory, with applications, Psychometrika, XXX, 1965, 239.

Lord, F.M., and Novick, M.R., Statistical theories of mental test scores, Reading, Mass: Addison-Wesley, 1968.

Panchapakesan, N., The simple logistic model and mental measurement, Doctoral dissertation, University of Chicago, 1969, (Unpublished).

Rasch, Georg, Probabilistic models for some intelligence and attainment tests, Copenhagen: Danish Institute for Educational Research, 1960.

Rasch, Georg, On general laws and the meaning of measurement in psychology, In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics, Berkely: University of California Press, 1961, IV, Pp. 321-334.

Rasch, Georg, Objective comparisons, In Lectures given at the UNESCO Seminar, Voksenasen, Oslo, 1964, (Unpublished).

Rasch, Georg, An individualistic approach to item analysis, In Readings in mathematical social science, Edited by P. Lazarsfeld and W. Henry, Chicago: Science Research Associates Inc., 1966, Pp. 89-107, (a).

Rasch, Georg, An item analysis which takes individual differences into account, British Journal of Mathematical and Statistical Psychology, XIX, Part 1, 1966, Pp. 49-57, (b).

Rentz, C.C., An investigation of the invariance properties of the Rasch Model parameter estimates, Doctoral dissertation, University of Georgia, 1975, (Unpublished).

Rentz, C.C., and Bashaw, W.L., Equating reading tests with the Rasch model. Final Report, Educational Research Laboratory, College of Education, University of Georgia, 1975.

Ross, J., An empirical study of a logistic mental test model, Psychometrika, XXXl, 1966, 325.

Sedlacek, W.E., Medical College Admission Test, Handbook for Admissions Committees, Second Edition, 1967.

Thorndike, E.L., et al., The Measurement of Intelligence, New York: Teacher's College, Columbia University, 1926.

Thurstone, L.L., A method of scaling psychological and educational tests, Journal of Educational Psychology, XVI, 1925, 433.

Vogt, D.K., An extension of the Rasch model to the case of polychotomously scored items, Doctoral dissertation, University of Maryland, 1971, (Unpublished).

Wright, B.D., Sample-free test calibration and person measurement, In Proceedings of the 1967 Invitational Conference on Testing Problems, Princeton: Educational Testing Service, 1968, Pp. 85-101.

Wright, B.D., Fit of the Rasch model to data, Presentation at the 1969 AERA Presession on Person-Free Item Calibration and Item-Free Person Measurement, Los Angeles, California, (Unpublished).

Wright, B.D., and Douglas, G.A., Best test design and self-tailored testing, Research Memorandum, No. 19, Statistical Laboratory, Department of Education, University of Chicago, 1975, (a).

Wright, B.D., and Douglas, G.A., Better procedures for sample-free item analysis, Research Memorandum, No. 20, Statistical Laboratory, Department of Education, University of Chicago, 1975, (b).

Wright, B.D., and Mead, R.J., CALFIT: Sample free calibration with a Rasch measurement model, Research Memorandum, No. 18, Statistical Laboratory, Department of Education, University of Chicago, 1975.

Wright, B., Panchapakesan, N., A procedure for sample-free item analysis, Educational and Psychological Measurement, XXIX, 1969, Pp. 23-48.