The application of measurement theory to tests in mathematics: a study of the goodness-of-fit of rasch model to the alis mathematics test

Siu Kam Kwan

THE APPLICATION OF MEASUREMENT THEORY TO TESTS IN MATHEMATICS: A STUDY OF THE GOODNESS-OF-FIT OF RASCH MODEL TO THE ALIS MATHEMATICS TEST

by

Siu Kam KWAN

Submitted to the School of Education

in partial fulfillment of

the Requirement for the degree of

Doctor of Education

University of Durham

2003

Table of Contents

Declaration

I declare that no part of the material in this thesis has previously been submitted by me for a degree in this or any other university.

Acknowledgement

I wish to express my gratitude to my supervisor, Professor Carol Fitz-Gibbon of the School of Education, University of Durham. From her, I not only acquired an appreciation of the measurement problem and her thoughtful guidance throughout my doctorial studies, but also her generosity to allow me to use the data of International Test of Developed Ability collected by the Curriculum, Evaluation and Management Centre. Gratitude is also extended to Professor Peter Tymms for his valuable suggestions on the research.

My thanks also go to Mr Peter P.L. Lee and Mr K.S. Leung, the senior curriculum officers of mathematics of the Curriculum Development Institute of Hong Kong for their assistance in sorting the items of the mathematics test studied in this research into different content areas.

Lastly, I am also indebted and grateful to Catherina, my wife for her understanding and encouragement, and my kids, Justin and Maria for their moral support.

Abstract

The scores provided by the International Test of Developed Ability (ITDA) have been used as an alternative baseline for comparing the progress of students in the A-level Information System (ALIS) project of U.K. The responses of 26,964 examinees to the mathematics items of ITDA in year 2000 were fitted by using the Rasch model. Five subject groups (the population, 2 gender groups and 2 ability groups) and 25 random samples (5 from each group) were generated from the responses of the examinees. The unconditional maximum likelihood estimates of the item difficulty and examinee ability parameters for various groups/samples were produced by the RASCAL program.

The scatterplots among different sets of sample item difficulty parameters reflected that the feature of item and ability invariance was not preserved in the groups of extreme abilities. The assumptions of unidimensionality, equal item discrimination, zero guessing factor and non-speededness were generally not supported in the two ability groups. In particular, the result indicated that the ITDA Mathematical Test might be a speeded test.

It was quite interesting in this study to see that the item difficulty parameters and examinee abilities estimated from the Classical Test Theory (CTT) and those from the Rasch model were very comparable and both frameworks exhibit more or less the same feature in terms of invariance. On the other hand, more items were "found" unfit by the CTT method than the Rasch approach indicating that the former looks more sensitive to the lack of fit than the latter.

To study the effect of speededness, the analysis was repeated with the last 11 items (which has the highest omits) deleted. Disappointingly, the results showed no significant improvement. Further research on the fitness of data with speed incorporated into the estimation of ability level is recommended.

CHAPTER 1

STATEMENT OF THE PROBLEM

Introduction

1.1 Concern about the quality of school education is an international trend. Educators and administrators make efforts to devise "indicators" measuring and monitoring school performance and value-added improvement in student performance in major domains of education. According to Willms (1992, p.1), an indicator is "simply a statistic describing some feature of the schooling system associated with its performance". Fitz-Gibbon (1996, p.5) defines an indicator as "an item of information collected at regular intervals to track the performance of a [schooling] system". Examples of indicators include the average test score of a school, the percentage of drop-outs, the civic awareness and moral attitudes of students, students' ability to cope with pressure and changes, etc. Clearly, a good indicator will provide good information for measuring performance and monitoring system. However, it is not easy to formulate a set of commonly acceptable, measurable and reliable indicators to assess the value-added performance of schools and students. The A-level Information System (ALIS) in U.K. is an up-and-running indicator system which has been providing "value-added" measures to schools and colleges over a decade.

1.2 The ALIS is a project which aims to provide performance indicators for post-16 students of U.K. across all sectors of education. It is a value-added monitoring

system run at the Curriculum, Evaluation and Management (CEM) Centre of the University of Durham. It has grown from only 12 schools in 1983[1] to cover one third of all A-Level entries in U.K. and international schools sitting U.K. examinations in 1999. The ALIS project has been extended to cover the Advanced General National Vocational Qualifications (GNVQs) examinations in the last few years and is hence sometimes called ALIS+. But the addition of vocational qualifications is now simply subsumed under the one name ALIS without the "plus".

1.3 The progress of students in the ALIS project is compared based on the performance indicators provided so that schools joining the project will know exactly the locations of their students. To make these comparisons fairly, all students are measured against a common baseline representing their ability before starting their post-16 courses. Generally, the GCSE scores are the baseline for the ALIS project. However, for students with no GCSE scores, the scores provided by the International Test of Developed Ability (ITDA) are used as an alternative baseline.

1.4 The ITDA was developed at the Educational Testing Service, Princeton, New Jersey of U.S.A. as a measure suitable for college entrants around the world and was introduced into the ALIS project in 1988[2]. The ITDA comprises a 20-minute verbal section of 25 questions, a 25-minute mathematical section of

[1] In 1983, the ALIS project was first introduced as Confidential, Measurement-Based Self-Evaluation (COMBSE) project. The existing name was used since 1988 (Fitz-Gibbon, 1996, pp.54-61).

[2] The AH6 SEM test of the National Foundation for Educational Research and the Raven's Advanced Progressive Matrices were used as the ability test in 1984 and 1985 respectively. Since these two tests were not good predictors of the A-level grades, the ability test was dropped until 1988 (Fitz-Gibbon, 1996, pp.58-61).

35 questions and a 5-minute vocabulary test (introduced in 1992 to improve the predictive validity of the test) of 20 questions. All the questions are multiple-choice items.

1.5 Since the introduction of ITDA to the ALIS project, the ITDA mathematics items were unchanged. Consequently, the test may have lost its effectiveness as an objective instrument for an alternative baseline for comparing the progress of all students in the ALIS project. Of course, a strict security may improve the situation but no test could be strictly secure if it was repeated year by year. Therefore, a viable alternative is to have different tests set for students in the ALIS project every year. The tests then not only provide a baseline for measuring the "value-added" of students but also let schools joining the project know the exact "locations" of their students. However, this alternative naturally gives rise to the problem of maintaining the standard of the test so that abilities of different examinees can be compared. A measurement model which can generate sample free item difficulties is very useful for such purpose. A pool can then be constructed with items fitted and calibrated by the measurement model and the mathematics items of the ability test could be drawn from the pool. This research intends to investigate the extent to which the ITDA mathematics items can be fitted by using measurement theory. Two theories, namely the *Classical Test Theory* (CTT) and the *Item Response Theory* (IRT) or *Latent Trait Theory* could be considered and they are discussed as follows.

Classical Test Theory

1.6 The CTT can be used to specify rules for transforming examinees' responses to the items of an educational test into estimates of their *latent abilities* or *traits*[3] ("latent" because the abilities or traits are not directly measurable) assumed to underlie the observable responses. The CTT model is simple. An examinee's score X on a particular test is considered to be a chance variable with some unknown frequency distribution. The mean of this distribution is the examinee's true score T. Mathematically,

$$X = T + E \qquad (1.1)$$

where E is the error of measurement. The basic assumptions on the error term in (1.1) are (a) the expected value of error over examinees is zero; (b) the error is not related to true score, other error scores and other true scores; and (c) the errors are normally distributed within examinees and homogeneously distributed across persons (e.g. Lord, 1980, pp.4-6; Embretson & Reise, 2000, pp.42-43).

1.7 The CTT mainly focuses on test-level information, but item statistics (i.e. item difficulty and item discrimination) are also essential in the model. The major advantage of the CTT is that it has relatively weak theoretical assumptions and hence can be easily applied in many testing situations (Fan, 1998, p.358). The estimation of the item parameters is simple and straightforward. The CTT collectively considers a group of examinees and examines their success rate on an item. This success rate, expressed as the proportion of examinees correctly respond to the item, is known as the *p-value* of the item and is used as the index

[3]According to Brown (1976, p.5), a trait can be considered as "a cluster of interrelated, or intercorrelated behaviours. ... It is an abstraction, a construct, rather than an objective, tangible reality."

for the *item difficulty* (the higher the p-value, the less difficult is the item). The *item discrimination* is expressed as the Pearson product-moment correlation coefficient between the scores on the item and the scores on the total test. When the items are dichotomously scored (i.e. 1 for a correct and 0 for an incorrect response), that coefficient is estimated as the point-biserial correlation coefficient. The purpose of examining the item-total correlations is to find out whether a particular item tends to be failed by examinees of low-ability and passed by examinees high in ability (reflected by high correlation) or vice versa (reflected by low correlation).

1.8 However, the CTT has also a lot of shortcomings (Hambleton & Swaminathan, 1985, pp.1-4; Hambleton, 1995, pp.84-85; Fan, 1998, p.358). Among all, the most important shortcoming is that the examinee characteristics and the test characteristics are inter-dependent and cannot be separated. For example, in the CTT, the ability of an examinee (which is the examinee characteristics we are interested in) is expressed by the true score (i.e. the expected value of observed performance on the test) and is hence only defined in terms of a particular test. The definitions of the item parameters (i.e. the item difficulty and item discrimination) make it clear that their estimates depend on the ability of examinees being measured. Therefore, the examinee parameters (i.e. ability) and the item parameters are not sample-free. Many measurement specialists discredit the CTT model with regard to its inability to produce examinee-invariant item parameters and item-invariant examinee parameters. The dependence of item measurement on a particular group and the dependence of examinee scores on a particular test are serious concerns as they lead to a lack of comparability of examinees who have taken tests of different difficulties and

tests taken by different groups of examinees.

Item Response Theory

1.9 IRT is another technique used in item analysis. The IRT approach is based on the postulate that for each item, there is a curve, known as *Item Characteristic Curve* (ICC), which relates the probability of getting an item right to the examinee's ability. Each item has its own curve, which is expected to move from left to right in an upward direction. Four variables are required to define an ICC or to determine the probability. There are one examinee parameter and three item parameters. The examinee parameter is the *ability* or *trait* and the three item parameters are the *item difficulty*, *item discrimination* and *guessing level*. Depending on the number of item parameters needed to describe the ICC, three basic models result. They are the one-, two- and three-parameter logistic models. Figure 1.1 shows three typical ICCs for the one-parameter logistic model. They are "parallel" curves which differ only by their location on the ability scale and the item difficulty is the only item characteristic that influences examinee performance. Figure 1.2 shows two typical ICCs for the two-parameter logistic model. The curves have different slopes reflecting that their item discrimination values are different. Four typical three-parameter logistic ICCs are shown in Figure 1.3. They have highlighted the role of guessing level in the shape of ICCs. Clearly, the non-zero probability of success for the low ability examinees indicates the existence of the guessing factor.

<Figure 1.1 – 1.3>

1.10 In the general IRT, it is assumed that a set of n latent abilities underlies examinee performance on a set of test items. The n latent abilities define a *n-dimensional latent space*. Each examinee's location in the latent space can be easily determined if the examinee's position on each latent ability is known. The latent space is referred to as "complete" (Lord & Novick, 1968; p.359; Hambleton & Swaminathan, 1985, p.16) if all the n latent abilities have been specified. In the item response model, it is commonly assumed that the latent space is *unidimensional*, that is, a single latent ability is adequate to specify an examinee's test performance. In fact, this assumption can hardly be met because factors like the examinee's personality, the examinee's ability to work quickly, test anxiety, level of motivation, etc. always exist and influence the test performance. Therefore, a more realistic approach is to seek "a 'dominant' component or factor that influences test performance" (Hambleton & Swaminathan, 1985, p.17). Clearly, it is this dominant component which is referred to as the ability measured by the test.

1.11 Apart from the assumption of unidimensionality, the one-parameter logistic model further assumes that all items are equally discriminating and there is no guessing. Therefore, only one item parameter – the item difficulty (or the location with respect to the ICC) is specified in this model. The appropriateness of these assumptions has been criticized by many measurement practitioners (e.g. Divgi, 1986). However, the assumptions can be found not unreasonable if one looks into the questions of how the word "harder" is interpreted and what makes examinees "guess" in the test. If it is agreed that the probability of success on the harder of two items should always be less than

that on the easier, then different item discriminations which will produce crossed ICCs does not occur. Therefore, variation in discrimination can be considered as "a symptom of item bias, multi-dimensionality" (Wright, 1992, p.199). On the other hand, if the items are too hard for an examinee, then he or she will be forced to guess. Therefore, if the difficulties of items "match" with the abilities of examinees, the guessing factor will be reduced to a minimum. On the whole, as commented by Hambleton et al (1991), the assumptions of equal discrimination and zero guessing can be approximately met "for relatively easy tests constructed from a homogeneous bank of test items" (pp.13-14).

1.12 There is also an implicit assumption of IRT models, namely, the non-speededness – the tests under the study of fit are not administered under speeded condition. If an examinee does not have time to reach an item, then his/her response does not only depend on his/her ability but also on his/her speed of performance. Since the unidimensional IRT models only deal with actual responses, the behaviour of examinees who do not respond to some of the items in a test because of lack of time cannot be described by the model. Moreover, apart from speed itself, psychological effect like nervousness, motivation, etc. resulting from speededness also have impact on the test performance. In general, speed is usually considered as a dimension independent of the trait measured by the test content. But since it affects mathematics achievement, it is sometimes confounded with the trait of examinees and is not easy to be identified. For example, when an examinee does not answer an item, the item may be too difficult to the examinee so that he/she just leaves it unanswered or he/she has no time to reach that item. On the other hand, having speed may be important for mathematics aptitude as examinees who can not only do

mathematics, but also do it quickly will tend to get higher grades than those who need more time.

1.13 Georg Rasch is the main proponent of the one-parameter logistic model, which is therefore commonly known as the *Rasch model.* As long as all the items in the bank had equal discrimination, it was possible to produce a single item parameter (the item difficulty) which was independent of the examinees and the other items in the test (Rasch, 1966, pp.53-56; 1980, pp.178-182). That is, the Rasch model is "person free" and "item free". Therefore, if the assumptions are met, the Rasch model can be used to calibrate a pool of items and any subset of items from the pool can be used to measure examinees on the same ability scale.

Source of Data

1.14 The data for the study consist of responses of 26,964 students to the 35 mathematics items of the ITDA offered in the ALIS project. The items are of multiple-choice type, each with 4 options. The data were collected by the CEM Centre in 2000 and had been scored. Each line of data in the data file consists of the identity and sex of the examinee and his/her scores (i.e. 0s or 1s) to the 35 items in the Mathematical Test. There are totally 26,964 lines in the set. Figure 1.4 shows part of the file. The first character represents the gender of the examinee concerned (1 for male and 2 for female), the next 7 characters are the examinee identification (e.g. 0121852 in the first line), the 9th character is a space and the sequence of 1s and 0s following (totally 35 numbers) are the scores of the individual items. For example, the examinee 0121852 has

got a correct response (i.e. score 1) to items 1 and 2, an incorrect response (i.e. score 0) to item 3, etc. "x" is a code for omitted responses and not-reached responses. It can be seen that some examinees had just left some items unresponded. For example, examinee 0291852 (the 4th line from the bottom in Figure 1.4) did not reach the last 20 items and hence left them unattempted.

<Figure 1.4>

1.15 The ITDA mathematics items are primarily curriculum content based. The frequency distribution and cumulative frequency distribution are respectively shown in Figure 1.5 and 1.6. It is not hard to see that the distribution is not normally distributed; rather it exhibits an obvious floor effect (skewness = 1.692), that is, the scores of most examinees are at the lower end of the distribution.

<Figure 1.5 – 1.6>

Aims of the Study

1.16 The CTT and IRT models, particularly the Rasch model, are widely perceived as two very different measurement frameworks. The major advantage of the CTT is its relatively weak theoretical assumptions so that it is easy to apply in many testing situations and these tests usually have quite acceptable internal consistency. On the other hand, the IRT solves problems by focusing on the interaction of persons with items. The item and person parameters estimated are sample independent which is a very crucial feature and a strong argument in favour of the IRT framework. It would be of great interest to compare the two

frameworks and see the extent to which what is learnt from each is much the same or different. Therefore, the aims of this study are to:

(a) determine the extent to which the Rasch model fits the data of the mathematics items of the ITDA;

(b) study the unidimensionality of the ITDA mathematics items based on the item parameter estimates;

(c) investigate to what extent the items the model fits satisfy the assumption of equal item discrimination;

(d) investigate to what extent the items the model fits satisfy the assumption of no guessing;

(e) investigate to what extent the items the model fits satisfy the assumption of non-speededness;

(f) compare the classical test theory method with the Rasch approach; and

(g) identify poor items using independent analysis from the classical test theory method and the Rasch approach.

1.17 In this study, it should be noted that omits and items not reached could not be easily distinguished because examinees might not respond to the items in serial order. Therefore, any attempt to assume that not-reached items only occur at the end of the test and ignore the not-reached items (say scoring them as zero) in the analysis would set limits on the conclusions that can be drawn from the study.

CHAPTER 2

METHODS REVIEW

Review of Item Response Theory

2.1 The IRT was developed over 50 years. Essential work embraces those of Tucker (1946) who was the first to use the term "item characteristic curve", a key concept in the field of IRT, Lord who described the two-parameter normal ogive model (1952), derived parameter estimates (1953a) and considered application of his model (1953b), Birnbaum (1968) who used the logistic curve for Lord's normal ogive curve and Rasch (1980) who introduced his probabilistic model for item analysis, etc. Thereafter, the important breakthrough in problem areas of test score equating, adaptive testing, test design and test evaluation by using IRT has attracted the interest of many measurement specialists (Hambleton & Swaminathan, 1985, p.7).

2.2 An essential part of IRT is using a mathematical function to relate the probability of giving a correct response (or the probability of success) to an item by an examinee to certain characteristics of the examinee and the item. Many different IRT models have been considered. Although whether these models provide satisfactory solutions to measurement problems can only be tested empirically, they have a common perspective, that is, they assume that the probability of an observable response is related to the examinee's latent ability. The ability is not directly observable, but the crucial point of IRT is to make use of the observed behaviour of the examinee (usually responses to a set of items

are represented as 0s or 1s) to estimate the examinee's ability. The latent ability is usually considered to be distributed continuously, but the particular form of the distribution is not known. The general class of measurement theories specifies how to transform the item responses into an estimate of the ability underlying the item response.

2.3 Some different IRT models have been used in the past. Examples include the Guttman "perfect scale", the linear model of Lazarsfeld (cited in Hulin et al, 1983, pp.16-19), the normal ogive model (Lord, 1952) and the logistic model (Birnbaum, 1968; Rasch, 1980).

(a) The Guttman "perfect scale" used the step function

$$P(\theta) = 0 \text{ for } \theta < b; \qquad (2.1)$$
$$= 1 \text{ for } \theta \geq b$$

where θ is the latent ability, to relate the probability of a response to an item with the ability under consideration. The discontinuity at the breaking point b of the ability and the flatness of the curve seem unrealistic as people rarely behave in this way.

(b) The linear model was developed by Lazarsfeld in which

$$P(\theta) = a + b\theta . \qquad (2.2)$$

One obvious limitation of this model is that the probability cannot lie between 0 and 1 unless b is zero, but in this case, the item provides no information about θ.

(c) The normal ogive model postulates a normal cumulative distribution function as a response function for the item i of a set of test items:

$$P(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \qquad (2.3)$$

where a_i is the item discrimination parameter of the item i and b_i is the corresponding item difficulty parameter. The normal ogive expresses the area under the standardized normal curve from a z-score of $-\infty$ to a z-score of $a_i(\theta - b_i)$ as the probability of success. It is viable as the said area is always positive and less than 1. However, a defect of this model is its assumption of normality of the distribution of the examinee ability (van der Linden & Hambleton, 1997, p.7).

2.4 Rasch developed the logistic model in 1950s. In this model, Rasch proposed

$$P(\Theta_x) = \frac{\Theta_x}{\Theta_x + B_i} \tag{2.4}$$

where Θ_x is the ability of an examinee x and B_i is the difficulty of item i. Clearly, this model describes the probability of a successful outcome of an examinee on an item as a function of only the examinee's ability and the item's difficulty. Taking the parameters on a logarithmic scale, that is, taking the transformation:

$$\Theta_x = e^{\theta_x} \quad \text{and} \quad B_i = e^{b_i},$$

(2.4) becomes

$$P(\theta_x) = \frac{e^{\theta_x}}{e^{\theta_x} + e^{b_i}}$$

$$= \frac{1}{1 + e^{-(\theta_x - b_i)}}. \tag{2.5}$$

This transformation has assigned an ability of negative infinity to a score of zero and an ability of positive infinity to a score of 100 percent. (2.5) can be rearranged to give

$$\theta_x - b_i = \log \frac{P(\theta_x)}{1 - P(\theta_x)}.$$

It is noted that the ability and item difficulty have been expressed in the same scale known as *log-odds scale.* The unit on the scale is called a *logit* which is

defined as the natural logarithm of the quotient of the log-odd for success and is "the distance along the line of the variable that increases the odds of observing the event specified in the measurement model by a factor of 2.718 ..., the value of e" (Linacre & Wright, 1989, p.54). Clearly, if log-odds are used in the vertical axis of the ICC in Figure 1.1 instead of the probabilities, the curves will become lines which are exactly parallel.

2.5 Since the Rasch model involves only one parameter – the item difficulty parameter, it is also known as one-parameter logistic (1PL) model. The logistic ogive is practically identical to the normal ogive as it has been shown by Haley that

$$\left| \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-z^2)dz - \frac{1}{1+e^{-1.7x}} \right| < 0.01 \qquad (2.6)$$

(cited by Birnbaum, 1968, p.399 and Van der Linden & Hambleton, 1997, p.13). Thus, the 1PL ogive is usually scaled as

$$P(\theta_x) = \frac{1}{1+e^{-D(\theta_x - b_i)}} \qquad (2.7)$$

where D = 1.7 to keep the inequality (2.6) true. Inequality (2.6) is important as it ensures that the logistic ogive and the normal ogive can approximate one another.

2.6 Birnbaum's main contribution to IRT was his suggestion to replace the normal ogive model by the 2PL model

$$P(\theta_x) = \frac{1}{1+e^{-Da_i(\theta_x - b_i)}}. \qquad (2.8)$$

He also proposed a third parameter – the pseudo-chance level or guessing level parameter to account for the non-zero performance of the low ability examinees on multiple-choice items. The 3PL model then takes the form

$$P(\theta_x) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_x - b_i)}}, \qquad (2.9)$$

where c_i is the guessing level parameter of item i. Clearly, the Rasch model is a special case of the 3PL model with unity a_i and zero c_i.

The Rasch Model and its Mathematics

2.7 The Rasch model has the following important features:

(a) The abilities[4] of people and difficulties of items are along the same scale so that abilities and difficulties can be compared. For a given item, a more able person is always more likely to get a correct response than a less able person; and a given person is always more likely to answer an easy item correctly than a difficult item. Here the term "more able person" is a shorthand for a person who is currently more able on the dimension being measured. Indeed, "ability" here often means a level of achievement.

(b) The Rasch model produces item difficulty levels independent of examinee samples and person abilities independent of the particular test administered.

(c) Sufficient statistics[5] exist. For example, all the information about the ability of a person on a given dimension is contained in the number of correct responses (or the number correct score). All persons with the same number correct score must be assigned the same estimated ability.

---

[4] The terms ability and achievement are sometimes contentious. Ability seems to often denote a fixed aptitude whereas achievement is altered by, for example, schooling.

[5] A sufficient statistic for a parameter Φ is a statistic that captures all the information about Φ contained in the sample concerned (Casella & Berger, 1990, p.247). That is to say, any additional information in the sample does not contain any more information about Φ.

Similarly, all the information about item difficulty is contained in the number of persons correctly responding to the item. All items with the same score must be assigned the same estimated item difficulty level.

(d) The model is a theoretical model and is relatively simpler than other logistic models. Its relative simplicity makes it less expensive and easier to apply in solving practical measurement problems. It is consistent with Kaiser's Principle II that "simplicity is elegance" (1970, p.404) and Einstein's view that "Everything should be as simple as possible but no simpler" (cited by Fitz-Gibbon, 2000).

The observable responses of a group of examinees to a set of items are usually obtained for a sample. The item and ability parameters are then determined by statistical estimation. Several estimation procedures are available, but in this study, the maximum likelihood procedure[6] is adopted.

2.8 Let the probability that an examinee x with ability θ_x obtains a response U_{xi} on item i with difficulty b_i be $P_{xi}(U_{xi}|\theta_x, b_i)$ where $U_{xi} = 1$ for a correct response and 0 for an incorrect response. For a correct response, the probability $P_{xi}(U_{xi} = 1|\theta_x, b_i)$ is the item response function and is commonly denoted as $P_{xi}(\theta_x, b_i)$ or simply P_{xi}. For an incorrect response, the probability is

$$P_{xi}(U_{xi} = 0|\theta_x, b_i) = 1 - P_{xi}(U_{xi} = 1|\theta_x, b_i) = 1 - P_{xi} = Q_{xi}.$$

Therefore, the probability of a response U_{xi} can be expressed as

$$P_{xi}(U_{xi}|\theta_x, b_i) = P_{xi}(U_{xi} = 1|\theta_x, b_i)^{U_{xi}} P_{xi}(U_{xi} = 0|\theta_x, b_i)^{1-U_{xi}}$$

[6] Maximum likelihood estimators are consistent and efficient estimators, functions of sufficient statistics and asymptotically normally distributed.

$$= P_{xi}^{U_{xi}} Q_{xi}^{1-U_{xi}}.$$

If the examinee x with ability θ_x responds to n items and if the assumption of unidimensionality holds, the joint probability of the responses $U_{x1}, U_{x2}, \ldots, U_{xn}$, denoted by $P_x(U_{x1}, U_{x2}, \ldots, U_{xn} | \theta_x, (b_i))$, where (b_i) is the 1 × n vector with elements b_i (i = 1, 2, 3, …, n), can be expressed as a product of the individual probabilities of the responses $U_{x1}, U_{x2}, \ldots, U_{xn}$. That is,

$$P_x(U_{x1}, U_{x2}, \ldots, U_{xn} | \theta_x, (b_i)) = \prod_{i=1}^{n} P_{xi}(U_{xi} | \theta_i, b_i)$$
$$= \prod_{i=1}^{n} P_{xi}^{U_{xi}} Q_{xi}^{1-U_{xi}}.$$

The likelihood function for the examinee x is therefore given by

$$L_x(u_{x1}, u_{x2}, \ldots, u_{xn} | \theta_x, (b_i)) = \prod_{i=1}^{n} P_{xi}^{u_{xi}} Q_{xi}^{1-u_{xi}} \qquad (2.10)$$

where $u_{x1}, u_{x2}, \ldots, u_{xn}$ are the respective specific values (either 0 or 1) taken by the random variables $U_{x1}, U_{x2}, \ldots, U_{xn}$.

2.9 Suppose a group of N examinees is administered a set of n items, the likelihood function from (2.10) is

$$L([u_{xi}] | (\theta_x), (b_i)) = \prod_{x=1}^{N} L_x(u_{x1}, u_{x2}, \ldots, u_{xn} | (\theta_x), (b_i))$$
$$= \prod_{x=1}^{N} \prod_{i=1}^{n} P_{xi}^{u_{xi}} Q_{xi}^{1-u_{xi}} \qquad (2.11)$$

where $[u_{xi}]$ is the N × n matrix with elements u_{xi} (x = 1, 2, 3, …, N; i = 1, 2, 3, …, n) and (θ_x) is the 1 × N vector with elements θ_x (x = 1, 2, 3, …, N).

The logarithm of the likelihood function is then given by

$$\log L([u_{xi}] | (\theta_x), (b_i)) = \sum_{x=1}^{N} \sum_{i=1}^{n} [u_{xi} \log P_{xi} + (1 - u_{xi}) \log Q_{xi}]$$

If μ represents θ_x or b_i, then

$$\frac{\partial \log L}{\partial \mu} = \sum_{x=1}^{N}\sum_{i=1}^{n}[u_{xi}\frac{P_{xi}'}{P_{xi}} - (1-u_{xi})\frac{P_{xi}'}{Q_{xi}}]$$

$$= \sum_{x=1}^{N}\sum_{i=1}^{n}\frac{(u_{xi}-P_{xi})P_{xi}'}{P_{xi}Q_{xi}} \qquad (2.12)$$

The maximum likelihood estimators of (θ_x) and (b_i) are obtained by solving the set of likelihood equations $\frac{\partial \log L}{\partial \mu} = 0$ (totally N + n equations).

2.10 For Rasch model, $P_{xi} = \frac{1}{1+e^{-D(\theta_x - b_i)}}$ and $Q_{xi} = \frac{e^{-D(\theta_x - b_i)}}{1+e^{-D(\theta_x - b_i)}}$.

$$\frac{\partial P_{xi}}{\partial \theta_x} = \frac{-1}{(1+e^{-D(\theta_x - b_i)})^2}\left[e^{-D(\theta_x - b_i)} \times (-D)\right]$$

$$= \frac{De^{-D(\theta_x - b_i)}}{(1+e^{-D(\theta_x - b_i)})^2}$$

$$= DP_{xi}Q_{xi}$$

Thus, $\frac{\frac{\partial P_{xi}}{\partial \theta_x}}{P_{xi}Q_{xi}} = D$.

(2.12) gives $\frac{\partial \log L}{\partial \theta_x} = D\sum_{i=1}^{n}(u_{xi} - P_{xi})$ for x = 1, 2, 3, ..., N

$$= D(r_x - \sum_{i=1}^{n}P_{xi}) \qquad (2.13)$$

where $r_x = \sum_{i=1}^{n}u_{xi}$, is the number correct score for the examinee x. The likelihood equations for the estimation of the abilities of the N examinees are

$$r_x - \sum_{i=1}^{n}P_{xi} = 0 \qquad \text{for x = 1, 2, 3, ... , N} \qquad (2.14)$$

Now, $\frac{\partial^2 \log L}{\partial \theta_x^{\,2}} = D\sum_{i=1}^{n}\left(-\frac{\partial P_{xi}}{\partial \theta_x}\right)$ for x = 1, 2, 3, ... , N

$$= -D^2\sum_{i=1}^{n}P_{xi}Q_{xi}. \qquad (2.15)$$

Clearly, $\frac{\partial^2 \log L}{\partial \theta_x^{\,2}} < 0$ as D^2, P_{xi} and Q_{xi} are all positive numbers. Thus,

the θ_x^* at which $r_x - \sum_{i=1}^{n} P_{xi} = 0$ gives a maximum value of the log L or equivalently the likelihood L and is hence the maximum likelihood estimator of θ_x.

2.11 Since the likelihood equations (2.14) are nonlinear, a numerical procedure, known as Newton-Raphson procedure, has to be used. In general, if x_m is the approximate solution for the equation f(x) = 0 at the m th stage, then the improved solution x_{m+1} is given by

$$x_{m+1} = x_m - \frac{f(x_m)}{f'(x_m)} \qquad (2.16)$$

(Hambleton & Swaminathan, 1985, pp.79-81; Stephenson, 1961, pp.347-351). The procedure is iterated until the difference between x_{m+1} and x_m, i.e. $x_{m+1} - x_m$, is less than a pre-established value, say 10^{-3}. When this happens, the process is said to have converged and x_m is taken as the approximate solution to the equation f(x) = 0. It can be shown that Newton-Raphson procedure converges rapidly provided that f'(x) is not zero or close to zero (Acton, 1970, pp.54-55).

2.12 For the Rasch model, the recurrence relation becomes

$$\theta_{m+1} = \theta_m - \frac{\dfrac{\partial \log L}{\partial \theta}}{\dfrac{\partial^2 \log L}{\partial \theta^2}} \qquad \text{for examinee x}$$

$$= \theta_m + \frac{r_x - \sum_{i=1}^{n} P_{xi}}{D\sum_{i=1}^{n} P_{xi}Q_{xi}} \qquad \text{(from (2.13) and (2.15))} \quad (2.17)$$

(The subscript for examinee x is dropped from θ for simplicity.)

2.13 Similarly, to estimate the item difficulty, we have

$$\frac{\partial P_{xi}}{\partial b_i} = \frac{-1}{(1+e^{-D(\theta_x - b_i)})^2}\left[e^{-D(\theta_x - b_i)} \times D\right]$$

$$= \frac{-De^{-D(\theta_x - b_i)}}{(1+e^{-D(\theta_x - b_i)})^2}$$

$$= -DP_{xi}Q_{xi}$$

$$\frac{\frac{\partial P_{xi}}{\partial b_i}}{P_{xi}Q_{xi}} = -D$$

and $$\frac{\partial \log L}{\partial b_i} = -D\sum_{x=1}^{N}(u_{xi} - P_{xi}) \qquad \text{for } i = 1, 2, 3, \ldots, n$$

$$= -D(s_i - \sum_{x=1}^{N} P_{xi}) \qquad (2.18)$$

where $s_i = \sum_{x=1}^{N} u_{xi}$ is the number of examinees who respond to item i correctly.

The likelihood equations for the item difficulty parameters b_i are

$$s_i - \sum_{x=1}^{N} P_{xi} = 0 \qquad \text{for } i = 1, 2, 3, \ldots, n \qquad (2.19)$$

$$\frac{\partial^2 \log L}{\partial b_i^{\,2}} = -D\sum_{x=1}^{N}\left(-\frac{\partial P_{xi}}{\partial b_i}\right) \qquad \text{for } i = 1, 2, 3, \ldots, n$$

$$= -D^2\sum_{x=1}^{N} P_{xi}Q_{xi} < 0$$

Thus, the b_i^* at which $s_i - \sum_{x=1}^{N} P_{xi} = 0$ gives a maximum value of the likelihood L and is the maximum likelihood estimator of b_i. Again, the equations (2.19) can be solved by the Newton-Raphson procedure. The relevant recurrence relation is

$$b_{m+1} = b_m - \frac{\frac{\partial \log L}{\partial b}}{\frac{\partial^2 \log L}{\partial b^2}} \qquad \text{for item i}$$

$$= b_m - \frac{s_i - \sum_{x=1}^{N} P_{xi}}{D\sum_{x=1}^{N} P_{xi}Q_{xi}}. \qquad (2.20)$$

(The subscript for item i is dropped from b for simplicity.)

2.14 It can be seen that when an examinee responds incorrectly to all the items, i.e. $r_x = 0$, equation (2.14) reduces to $\sum_{i=1}^{n} P_{xi} = 0$. Since P_{xi} is the probability of a correct response, this equation is satisfied only when $P_{xi} = 0$ for all i, i.e. when $\theta = -\infty$. Similarly, when an examinee responds correctly to all the items, i.e. $r_x = n$, the likelihood equation reduces to $\sum_{i=1}^{n} P_{xi} = n$ which is satisfied only when $P_{xi} = 1$ for all i, i.e. when $\theta = +\infty$. Clearly, maximum likelihood estimators do not exist for these cases. A viable way to address the problems of zero-correct scores and perfect scores is to eliminate the examinees concerned from the estimation procedure. In a similar way, the items to which no examinee responds correctly and the items to which all examinees respond correctly will be eliminated.

2.15 Since the number correct score is a sufficient statistic (see Appendix 1 for proof) for estimating the ability of an examinee, any examinee who gets a certain score will be estimated to have the ability associated with that score. Hence all examinees who get the same score will be estimated to have the same ability. The likelihood equations (2.14) and (2.19) can then be simplified to

$$k - \sum_{i=1}^{n} P_{ki} = 0 \qquad \text{for } k = 1, 2, 3, \ldots, n-1 \qquad (2.21)$$

$$s_i - \sum_{k=1}^{n-1} n_k P_{ki} = 0 \qquad \text{for } i = 1, 2, 3, \ldots, n \qquad (2.22)$$

where n_k is the number of examinees in score group k (i.e. with score k).

Note that k = 0 and k = n have already been excluded from the estimation procedure. The corresponding recurrence relations in the Newton-Raphson procedure are

$$\theta_{m+1} = \theta_m + \frac{k - \sum_{i=1}^{n} P_{ki}}{D\sum_{i=1}^{n} P_{ki}Q_{ki}} \quad \text{for each score group} \qquad (2.23)$$

and

$$b_{m+1} = b_m - \frac{s_i - \sum_{k=1}^{n-1} n_k P_{ki}}{D\sum_{k=1}^{n-1} n_k P_{ki}Q_{ki}} \quad \text{for each item.} \qquad (2.24)$$

2.16 In summary, maximum likelihood estimators are found from the roots of the likelihood equations which set the derivatives of the log likelihood equal to zero. To estimate the parameters, the logarithm of the ratio of number correct score to the number incorrect score for each score group k, i.e. $\log\frac{k}{n-k}$, is used as an initial estimate of the ability (Hambleton et. al., 1991, p.42). Treating the ability values as known, the parameter b_i^* for item i is estimated by solving just one equation out of (2.24). When the item parameters are known, the ability estimate θ_k^* for score group k is found from just one equation out of (2.23). This suggests an iterative procedure when the trial values θ_k^* (k = 1, 2, 3, ... , n − 1) are treated as known while solving (2.24) for the estimates b_i^* (i = 1, 2, 3, ... , n), then treat all item parameters b_i^* (i = 1, 2, 3, ... , n) as known while solving (2.23) for all trial values θ_k^* (k = 1, 2, 3, ... , n − 1). This process is repeated until the numerical values converge. This procedure which uses provisional ability estimates as known values, improves the values by using

subsequently estimated item parameters (which are also successfully improved) and finally provides point estimates for all item and examinee parameters is known as *joint* or *unconditional maximum likelihood procedure.*

Goodness-of-Fit Tests

2.17 Item and examinee parameters will be "person independent" and "item independent" respectively when the Rasch model and the data are fitted. Model data fit can therefore be assessed by checking the invariance of item difficulties and examinee abilities estimated by the Rasch model (Gustafsson, 1980, p.209; Hambleton et al, 1991, p.24). Although invariance of item difficulties and examinee abilities can never be observed in the strict sense, the "degree" to which it holds can be assessed when samples of test data are used. For the invariance of item parameters, one possible way is to draw two samples of different abilities from the population and estimate the item parameters in each sample. The correlation coefficient and the scatterplot of the difficulty values for the items of the two samples can be studied. A low correlation coefficient or a large amount of scatter observed may indicate a lack of invariance although it may be caused by model-data misfit, poor item parameter estimation or by the fact that one or more of the underlying assumptions may not hold for the data set. It should be noted that a high correlation coefficient alone is not sufficient to indicate a model-data fit. For a real model-data fit, the points in the scatterplot of the item difficulty values estimated from two samples should fall along the baseline "y = x". Similarly, ability estimates can be compared for different item sets (e.g. set of odd-numbered items vs. set of even-numbered items, set of easy items vs. set of hard items and sets of items

reflecting differing content categories). Invariance is established if the estimates are highly correlated and the points are closed to the baseline.

2.18 Apart from the scatterplot mentioned in para.2.17, the fit between the Rasch model and an examinee's responses can also be assessed by examining the item residuals which are the differences between the model's probabilities of correct responses and the examinees' observed item performance. In particular, the residual of examinee x with ability θ_x on item i is

$$y_{xi} = u_{xi} - P_{xi}$$

where u_{xi} is 1 or 0 and P_{xi} is the probability of a correct response. Since the number correct score is a sufficient statistic, the residual could be rewritten as

$$y_{ki} = P^*_{ki} - E(P^*_{ki})$$

where y_{ki} is the residual of examinees in score group k on item i, P^*_{ki} is the observed proportion of correct responses in score group k on item i and $E(P^*_{ki})$ is the expected proportion of correct responses using Rasch model. To take into account the sampling error of $E(P^*_{ki})$, the standardized residuals are usually used.

$$y_{ki} = \frac{P^*_{ki} - E(P^*_{ki})}{\sqrt{Var(P^*_{ki})}}$$

where $Var(P^*_{ki})$ is the variance of P^*_{ki}. Clearly,

$$E(P^*_{ki}) = P_{ki}$$

and $$Var(P^*_{ki}) = \frac{E(P^*_{ki})[1 - E(P^*_{ki})]}{n_k} = \frac{P_{ki}(1 - P_{ki})}{n_k}$$

where n_k is the number of examinees in score group k.

Thus, $$y_{ki} = \frac{P^*_{ki} - P_{ki}}{\sqrt{\frac{P_{ki}(1-P_{ki})}{n_k}}} \qquad (2.25)$$

The sum of squares of the standardized residuals of an item over the m score groups with $n_k \neq 0$ is distributed as a chi-square statistic which can be used as a measure of the fitness of the item.

2.19 The residual difference between observed and expected scores is the basis of fit analysis and all chi-square measures used adopt more or less the same rationale. For example, Yen (1981, p.246) rank ordered the examinees on the basis of their ability estimates and divided into 10 cells with approximately equal number of examinees per cell and defined the Q1 statistic for item i as

$$Q_{1i} = \sum_{j=1}^{10} y_{ji}^2$$

$$= \sum_{j=1}^{10} \frac{n_j(P^*_{ji} - P_{ji})^2}{P_{ji}(1-P_{ji})} \qquad (2.26)$$

where n_j is the number of examinees in cell j, P^*_{ji} is the observed proportion of correct responses in cell j on item i and P_{ji} is the mean of the probability of correct responses of the examinees in cell j. The degree of freedom is 10 – 1 = 9. On the other hand, Wright & Panchapakesan (1969, pp.44-46) essentially involves calculating the following residual

$$y_{ki} = \frac{f_{ki} - E(f_{ki})}{\sqrt{Var(f_{ki})}}$$

or $$y_{ki} = \frac{f_{ki} - n_k P_{ki}}{\sqrt{n_k P_{ki}(1-P_{ki})}} \qquad (2.27)$$

where f_{ki} represents the number of examinees in score group k answering item

i correctly. Two other fit statistics are also reported. They are the infit and outfit mean square statistics (Rasch, 1980, p.193-194; Linacre & Wright, 1994, p.360; Bond & Fox, 2001, p.176). The outfit is based on the conventional sum of squared standardized residuals. It is an outlier sensitive fit statistic that picks up rare events that have occurred in an unexpected way. On the other hand, the infit is an information-weighted[7] sum of the squared standardized residuals. It focuses on the overall performance of an item or person. To calculate infit, each squared standardized residual is weighted by its variance, summed and then divided by the sum of variance (Linacre & Wright, 1994, p.360; Bond & Fox, 2001, p.176). The result is

$$\text{Infit} = \frac{\sum_{\substack{k=1 \\ n_k \neq 0}}^{n-1} \text{Var}(P_{ki}^*) y_{ki}^2}{\sum_{\substack{k=1 \\ n_k \neq 0}}^{n-1} \text{Var}(P_{ki}^*)}$$

$$= \frac{\sum_{\substack{k=1 \\ n_k \neq 0}}^{n-1} (P_{ki}^* - P_{ki})^2}{\sum_{\substack{k=1 \\ n_k \neq 0}}^{n-1} \frac{P_{ki}(1-P_{ki})}{n_k}} \quad \text{(from (2.25))}$$

2.20 One of the problems of using the chi-square statistic is its sensitivity to sample size. The chi-square value may become significant owing to large sample size

---

[7] The information in an observation quantifies how much could be learnt about the modeled variable from that observation (Wright, 1996, p.504). For the Rasch model, the item information I_{xi}, in the dichotomous responses (i.e. 1 or 0) of examinee x to item i is given by $I_{xi} = D^2 P_{xi}(1 - P_{xi})$ where D is 1.7 and P_{xi} is the probability that the examinee x obtains a correct response to item i. The amount of information provided by item i at the ability level θ_x is inversely related to the standard error associated with the ability estimate θ_x.

(Hambleton & Swaminathan, 1985, p.153; Embretson & Reise, 2000, p.235). Hambleton & Murray (cited by Hambleton & Swaminathan, 1985, p.155) has conducted a simulation study to illustrate the problem associated with examinee sample size. The results show that when the sample size was increased from 150 to 2400, the number of misfitted items was increased from 20 to 42 (out of 50) at the 0.05 level! In fact, the dependence of statistical significance on sample size is not uncommon. The larger the sample size, the more likely to get statistically significant results (Carver, 1978, p.387). Therefore, apart from statistical significance, the magnitude of the misfit and the interpretation of its educational significance should be sought as these are the most important pieces of information to arise from a study. The educational significance of the misfit requires a judgment made in terms of the outcome measure (Fitz-Gibbon, 1984, p.136).

Review of Approaches for Assessing the Assumptions of Rasch Model

2.21 Whether there is a model-data fit between the Rasch model and a data set is based on strong assumptions, i.e. unidimensionality, equal item discrimination, zero guessing level and non-speededness. Undoubtedly, these assumptions cannot be met completely by any set of test data. Therefore, it is more appropriate to locate and interpret the departures of the Rasch model from the assumptions in practical measurement problems. Since one of the most critical assumptions of latent trait models is that a set of items should measure only one ability, testing the assumption of unidimensionality takes precedence over others in most related research studies.

Unidimensionality

2.22 Unidimensionality is a qualitative concept and no test can ever be perfectly unidimensional. There are various methods/indices for assessing unidimensionality, but many of them lack a rationale and some are just adjustments of others to take into account the criticisms (Hattie, 1985, p.139). Hambleton & Rovinelli (1986) also criticized that most of the methods are "only loosely connected to the various definitions in the psychometric literature" (p.287). Nevertheless, indices based on reliability (such as coefficient alpha), factor analysis, principal components, residual analysis and Bejar's analysis method[8] (1980) are the more popularly used or promising methods in the field.

2.23 Indices based on Reliability

(a) The coefficient alpha (or Kuder-Richardson Formula 20[9]) is a popularly used index which is based on reliability and is a measure of internal consistency of a test (i.e. the extent to which scores on each item and scores on other items of the test are related (Bartram, 1990, p.71)). It is the expected value of the correlation between any two random samples of items drawn from a pool like the given test. Mathematically, the alpha, r_{xx} is given by

$$r_{xx} = \frac{k}{k-1} \times \left[1 - \frac{\sum S_i^2}{S_T^2} \right] \qquad (2.28)$$

where k is the number of "parallel" parts of a test, $\sum S_i^2$ is the sum of the

[8] The rationale of Bejar's analysis method is that if the items in a test are unidimensional, then the grouping of the items for calibration will be irrelevant and the parameter estimates for items calibrated with different subsets of items should be identical aside from sampling errors.

[9] The coefficient alpha is applicable to multichotomous items while Kuder-Richardson Formula 20 is

part variances and S_T^2 is the variance of the total test score. If a test is regarded as comprising "k" parallel items, coefficient alpha can be considered as the theoretical average of all possible split-half reliability coefficients[10] of the test (Green, 1991, p.29). Since the coefficient alpha is defined in terms of the concept of "parallel tests" (whether with only one item or sets of items) and one of the assumptions of "parallel tests" is that the items measure a single factor or ability (Bartram, 1990, p.71), a high alpha is believed to be an index of "common factor concentration" (Cronbach, 1951, p.331) and shows that the items are "all related to the same underlying construct" (Green, 1991, p.28). It is this interpretation which relates alpha to dimensionality. However, Green et al (1977) have given a numerical example to show that it is possible to obtain a high value of alpha for a multi-dimensional test. They also noted that the value of alpha is dependent on test length and group hetergeneity[11] (also Green, 1991, p.28; Thorndike, 1997, pp.110-114). This contradicts the basic conception that the unidimensionality of a test should be independent of its length and the group of examinees.

(b) The concept of reliability could be extended to the broader and more flexible notion of generalizability which recognizes multiple sources of measurement error, estimates each source separately and hence optimizes the reliability (Shavelson et al, 1989, p.923). The main distinguishing feature of generalizability analyses is that they explicitly consider a group

---

more specific and is applicable to dichotomous items.

[10] Split-half reliability coefficient is a form of internal consistency that preceded coefficient alpha. It is calculated by correlating one half of a test with the second half. But a difficulty with this formulation is how to divide the test as different divisions resulted in different split-half coefficients.

[11] It is true as alpha is a measure of internal consistency which is always higher for longer tests (with greater potential variability of scores) and more diverse groups (orderly relationships are more clearly visible for more striking differences among examinees), i.e. alpha is a sample dependent index.

of items as a random sample from a universe of items (Goldstein & Wood, 1989, p.148). The generalizability theory (G Theory) provides a generalizability coefficient (G coefficient) which is similar to the reliability coefficient in CTT. The G coefficient reflects the proportion of variability in examinees' scores that is attributable to variability in the "universe score" (which is analogous to the true score in CTT), that is, variability in the examinees' knowledge, skills and so on (Shavelson & Webb, 1991, p.14). In a test like the mathematics section of the ITDA, there are four different sources of variability. They are those which arise from the differences in examinees' abilities, the differences in the difficulty of test items, the interaction between examinees and items and the randomness. The last two sources of variability cannot be disentangled and are usually lumped together as a residual. The G theory uses the analysis of variance (ANOVA) to partition the total variability among item scores into the effects for examinees, items and the residual. Once the variability among scores has been partitioned, the G coefficient is given by dividing the estimated examinee variance component by an estimated observed score variance. As reliability, a high G coefficient (which is in the range of 0 – 1) reflects the unidimensionality of a test.

(c) The confusing concepts of unidimensionality, reliability, internal consistency and homogeneity (used specifically to refer to the similarity of the item inter-correlations) have introduced problems in assessing indices of unidimensionality (Mc Donald, 1981, p.103 & 110; Hattie, 1985, p.157). Reliability (and also G coefficient), internal consistency and homogeneity are comparative concepts but not unidimensionality. We can say one test is more reliable, more internally consistent or more homogeneous than the

other test, but as Mc Donald (1981, p.103) indicates, unidimensionality is "an integer-valued quantity" and we cannot say one test is more unidimensional than the other. In fact, a unidimensional test may or may not be necessarily reliable, internally consistent or homogeneous.

2.24 Factor analysis[12]

(a) McDonald (1981, p.101; 1982, p.394) defined a set of test items as unidimensional if the residual covariation between items in the set is zero for examinees with the same ability and argued that the dimensionality of a set of test items should be determined by the number of factors needed for describing examinees. Therefore, if only one factor is defined, the set of items should be unidimensional. However, the use of linear factor analysis is not appropriate since it assumes linearly related variables which contradict the non-linearity assumption of latent trait models. A major drawback of using factor analysis on dichotomous items is the distortion of the loadings of the items on the two extremes of difficulty scale (Hattie, 1985, p.149). This makes it appear that such items do not measure the same underlying dimension as the others, hence leading to a factor solution with too many factors. The occurrence of these "spurious factors" in addition to "genuine factors" is common in factor analyses when items of different difficulty levels are scored dichotomously and this raises problems in using factor analysis. These spurious factors are referred to as

[12] The basic notion of factor analysis is that the variance of an item is composed of 2 components, namely the common variance and the unique variance. The former is the proportion of the total variance that an item shares with the other items in the analysis while the latter is the remainder of the total variance in each item which consists both of a variance that is specific to a particular item and random error variance. The objective of factor analysis is to identify those factors arising only from the components of common variance of the set of items.

"difficulty factors" by a number of writers like Wherry & Gaylord and Carroll (cited in Hattie, 1985, pp.148-149) and they had been attributed to the attenuation of the phi coefficient[13] below what it should be if the difficulty levels were all the same. Wherry & Gaylord (cited in Hattie, 1985, pp.148-149) argued that phi correlations are contingent upon difficulty and hence suggested to use tetrachoric correlations[14] which would be 1.0 provided that the items measure the same ability regardless of differences in difficulty. However, the tetrachoric correlations need a normality assumption which is not a necessity to the Rasch model (Lord & Novick, 1968, pp.345-346; Slinde & Linn, 1979, p.440). Also, the normality assumption is hard to hold in multiple-choice tests because of the effect of guessing (Slinde & Linn, 1979, p.440). The calculation of tetrachoric correlations is complicated and does not necessarily yield a correlation matrix which is positive definite (Hambleton & Swaminathan, 1985, p.156). Mc Donald & Ahlawat (1974, p.98) opined that the occurrence of spurious factors is not due to the attenuation of the phi coefficient. They showed that they are due to the non-linear regressions of items on the factors rather than difficulty per se. They concluded in their paper that the notion of difficulty factors should be dropped and replaced by that of "factors due to non-linearity" (p.98). On the whole, no rigorous theory of "difficulty factors" seems to exist.

(b) Non-linear factor analysis may be more appropriate as it does not require the assumption of linear relationships among the variables and between the

[13] A phi correlation is a special case of Pearson correlation coefficient. It is a measure of the relationship between two dichotomous variables instead of continuous variables.

[14] A tetrachoric correlation is similar to the phi correlation, but it further assumes that there is a hypothesized and normally distributed continuous variable underlying each dichotomous variable.

variables and the abilities. However, non-linear factor analysis is not without problems. There are so many non-linear curves and it is hard to determine which is more appropriate. One popular version is to use a polynomial (in fact, any curve can be approximated by a polynomial), but the problem that naturally arises is the number of terms of the polynomial that should be retained. The procedure will be very complicated if too many terms in the polynomial are retained in the solution. Again, no accepted criterion for the appropriate number of factors to be retained in a solution is available (Hambleton & Rovinelli, 1986, p.301).

2.25 Principal components[15]

The first principal component which explains the maximum variance is used as an index of unidimensionality. The rationale is that the larger the amount of variance explained by the first component, the more likely the set of items to be unidimensional. However, there is no established criterion for "how high" the variance needs to be. For example, Carmines & Zeller (1979, p.60) recommended that at least 40% of the variance should be explained by the first component while Reckase (1979, p.227) recommended 20%. Moreover, it is not hard to show that a multi-dimensional set of items can have higher variance on the first component than does a unidimensional test (Mc Donald, 1981, p.112; Hattie, 1985, p.146). Therefore, the proportions of variance due to the first principal component are only "crude and unsatisfactory criteria for unidimensionality" (Mc Donald, 1981, p.113). A plot of eigenvalues of the

[15] Principal component analysis differs from factor analysis in that it is concerned with the total variance of an item. There is no distinction between common variance and unique variance. All variances are treated as common variances. In principal component analysis, only the first few components which account for most of the variance are retained for analysis.

inter-item correlation matrix (looking for a dominant first factor) and a comparison of two eigenvalue plots, one from the inter-item correlation matrix using the test data and one from that of a set of random data, have also been used. A common problem of using this approach is again how large is the eigenvalue to be retained and considered as a "dominant" factor. Since many of the components may not be interpretable or have large error variance, Kaiser (1970, p.408) recommended to retain those components with eigenvalues greater than one. Therefore, the number of eigenvalues greater than one can be used as an index of dimensionality. Lumsden (1961) suggested to use the ratio of the first and second eigenvalues as the index of unidimensionality. Hutten (cited by Hattie, 1985, p.146), also used the ratio of the first and second largest eigenvalues of matrices of tetrachoric correlations to assess unidimensionality and considered that high values of the ratio should indicate unidimensional tests while low values indicate multi-dimensionality. Lord (1980, p.21) argued that if the first root is large compared to the second and the second is not much larger than any of the others, then the set of items is roughly unidimensional. However, Hattie (1985, p.146) has shown that it is not necessarily true. The sum of squared residuals (or sum of the absolute values of the residuals) after removing one component has also been used, but "there is no established criterion for how small the residual should be" (p.146).

2.26 Residual analysis

The residual analysis involves fitting a unidimensional IRT model to the test data, using the model parameter estimates to predict the item performance data. The predicted values are then compared with the actual values by considering the resulting residuals. Again, no criterion is established for the size of the

residuals so that unidimensionality can be assumed. Moreover, large residuals may be due to the violations of model assumptions rather than unidimensionality (Hambleton & Rovinelli, 1986, p.300). This method in the study of Hambleton & Rovinelli (1986) was proved to be disappointing.

2.27 Bejar's analysis method

Bejar's method (1980) is based on the rationale that the grouping of the test items of a unidimensional test should be irrelevant in item calibration. That is to say, the performance of examinees on any subset of the test should be the same. The procedure requires to sort a priori the test items into categories, each of which appears to measure different traits. Then conduct a logistic model analysis of only the items in each category and repeat the model analysis using the total set of items. Two sets of item parameter estimates obtained for each item can be compared by using a statistical test or by plotting the estimates to determine the extent to which the two sets are linearly related. The merits of the Bejar's analysis method are that it does not involve linearity assumptions about the test data and it provides a straightforward check on the unidimensionality of the set of test data (Hambleton & Rovinelli, 1986, p.288). This method is particularly useful for achievement tests since the items frequently cover different content areas. In this case, it may be postulated that in addition to the "general" ability, a "unique" ability is being measured by the items within each content area. However, there is a drawback in this method: the number of items available in each category of items identified may be too small[16] for calibration (Bejar, 1980, p.294).

[16] If the sample size is small, say 100 persons, tests of more than 20 or 30 items are needed to protect measurement from unacceptable disturbance (Wright, 1977, p.106).

Equal Item Discrimination

2.28 Only descriptive methods are available for checking this assumption (Hambleton & Swaminathan, 1985, pp.159-160). A viable check is analyzing the distribution of the item-test score correlations, say biserial or point biserial correlations[17]. If the distribution is "reasonably homogeneous", the assumption may be right (Hambleton et al, 1991, p.56).

Zero Guessing Level

2.29 This assumption can be checked by analyzing the performance of low-ability examinees on the most difficult items (Hambleton et al, 1991, p.57). Zero guessing can be assumed if the performance levels are close to zero. According to Baker (1964, 1965), the viability of the zero guessing assumption can be examined by investigating the item-test score plots. Again, the assumption is supported if the item performance for low-scoring examinees is close to zero.

Non-Speededness

2.30 Again, only descriptive methods are available for checking this assumption. The following percentages can be studied:

(a) Percentage of omits of each item

(b) Percentage of examinees completing the whole test

(c) Percentage completing 75% of the test

(d) Percentage of examinees who do not respond to the last few items of the test

[17] The point biserial correlation is the product moment correlation between a continuous variable and a dichotomous variable while biserial correlation further assumes that there is a hypothesized and normally distributed continuous variable underlying the dichotomous variable.

CHAPTER 3

METHODOLOGY

Introduction

3.1 In this study, the Rasch model was applied to analyze the data and the RASCAL program[18] of Assessment Systems Corporation was used to compute the estimates of the difficulty parameters of the items and the abilities of the examinees selected. The estimates obtained from this program are unconditional maximum likelihood estimates. On the whole, this study is divided into seven sections as follows:

(a) The fit of Rasch model to the ALIS data

(b) The Check of the Assumption of Unidimensionality

(c) The Check of the Assumption of Equal Item Discrimination

(d) The Check of the Assumption of No Guessing

(e) The Check of the Assumption of Non-Speededness

(f) The Comparison between the Classical Test Theory Method and the Rasch Approach

(g) Identifying Poor Items using Independent Analyses from the Classical Test Theory Method and the Rasch Approach

[18] RASCAL is a one-parameter logistic model item calibration and test scoring program. It uses the unconditional maximum likelihood calibration method to estimate item difficulty parameters. Correction for the bias is an option. The scaling factor D can be set to 1.7 to produce results comparable to those produced by programs such as LOGIST or BILOG, or set to 1 to produce results on similar scales to those obtained with programs such as BICAL or BIGSTEPS. In this study, D is set to 1.7 and the ability scores are scaled to a mean of 0 and a standard deviation of 1.

Data Sets of the Study

3.2 Samples which are progressively more dissimilar were generated by 3 sampling plans so that the behaviours of statistics estimated by the Rasch model could be examined under different examinee sampling conditions. All samples generated in this way have a sample size of 1,000, which is considered sufficiently large for the estimation of IRT parameters.

(a) *Random samples* – Five random samples (named as r_sample1, r_sample2, etc.) were drawn from the whole data set of 26,964 examinees (named as Whole Group).

(b) *Gender group samples* – The Whole Group was separated into two gender groups, namely the Male Group (of 12,567 examinees) and the Female Group (of 14,397 examinees). Five samples of male examinees (named as m_sample1, m_sample2, etc.) and 5 samples of female examinees (named as f_sample1, f_sample2, etc.) were randomly drawn from each group respectively. As the male and female samples were drawn from different populations as defined by the gender variable, theoretically there should be more dissimilarity between a male sample and a female sample than between any two random samples described in (a).

(c) *Ability group samples* – Different samples in terms of performance on the ITDA Mathematical Test were generated in this sampling plan. The top 25% of examinees were arbitrarily labeled as High-ability Group and the bottom 25%, Low-ability Group. By examining the raw total scores which were obtained by first running the RASCAL program to the data in the Whole Group, it was found that examinees in the High-ability

Group (of 6,601 examinees) have raw total scores ≥ 18 (i.e. from 18 to 35) while those in the Low-ability Group (of 6,821 examinees) have raw total scores ≤ 10 (i.e. from 0 to 10). Five random samples were drawn from each of the ability groups (named as h_sample1, h_sample2, ... L_sample1, L_sample2, etc.). Since these two groups are defined in terms of test performance, not in terms of the gender variable as in (b), there should be more dissimilarity between a high ability sample and a low ability sample than between a male and female sample pair.

These 3 sampling plans totally generate 5 × 5 = 25 random samples. The data sets available for analysis in the study are summarized in Table 3:1.

<Table 3:1>

The Fit of Rasch Model to the ALIS Data

3.3 The research questions in this section are:

(a) How invariant are the item difficulty estimates across different participant samples?

(b) How well does the Rasch model fit item i of the Test (i = 1, 2, 3, ... , 35)?

(c) How invariant are the examinee ability estimates across different item sets?

3.4 Since the fit of the model is equivalent in concept to the invariance of item and examinee parameters, the invariance property was investigated in this section. To study the invariance of the item difficulty parameters, the RASCAL program

was applied to each of the 25 samples to obtain 25 sets of item difficulty parameter estimates. Correlation coefficients between two sets of difficulty values were computed as follows:

(a) Any two sets of item difficulty parameters from the 5 random samples were correlated. Totally, $C_2^5 = 10$ correlation coefficients were computed.

(b) Item difficulty parameters from each of the 5 random samples of male examinees were compared with those from each sample of female examinees resulting in $5 \times 5 = 25$ correlation coefficients.

(c) Similarly, item difficulty parameters from each of the 5 random samples of High-ability Group were compared with those from each sample of Low-ability Group resulting in another 25 correlation coefficients.

Since the sampling distribution of correlation coefficient is skewed (Hopkins, et al, 1996, p.260), the individual correlation coefficients obtained in (a), (b) and (c) were transformed into normal variates by the Fisher's transformation. The average value was then obtained in each group of samples (i.e. the random samples, the male samples, etc.) in the usual way. The average value was re-transformed back into the original scale to give an overall view of invariance. The correlation coefficient alone is not sufficient to test a model-data fit. Scatterplots of the difficulty estimates of some selected samples were obtained to examine whether the points fall along the baseline. Since the chi-square statistic is sample size dependent, it is pretty useless with the subject groups because of the large sizes. Instead, the chi-square statistics (χ_i^2, i = 1, 2, 3, ... , 35) for the 25 samples were computed for cross checking.

3.5 If there is a model-data fit, the item difficulty parameters estimated from the Whole Group should be the same as those from the gender and ability groups as the item parameters should be "person free". Therefore, the "effect" of the gender and ability of examinees on the magnitudes of misfit of the items were compared by studying the effect sizes of individual items. The relevant formula (Glass et. al., 1981, p.29; Fitz-Gibbon, 1984, p.138) is

$$\text{Effect size} = \frac{(\text{Experimental group mean}) - (\text{Control group mean})}{\text{Control group SD}}$$

In this study, the Whole Group (which contains all examinees) is taken as the control group and the item difficulty estimates of various subject groups were compared with the corresponding values estimated in the Whole Group by computing the "effect sizes" as follows:

$$\text{Effect size} = \frac{(\text{Subject group item estimate}) - (\text{Whole Group item estimate})}{\text{SD of the Whole Group item estimates}}$$

3.6 To study the invariance of examinee parameters, the test items were regrouped in three different ways to generate three pairs of item groups which were progressively dissimilar. The three groups are:

(a) *Equivalent-halves Groups* – The items were separated into two groups, one group of odd-numbered items and one group of even-numbered items.

(b) *Content Groups* – The items were separated into two groups according to different content categories (see para.3.7 below).

(c) *Difficulty Groups* – The items were divided into two groups based on the

item difficulty. Items with negative item difficulty levels were classified as easy items and those with positive item difficulty levels as hard items.

Six sets of ability estimates of the Whole Group were obtained for the 3 pairs of item groups by using the RASCAL program. The correlation coefficient between two sets of ability estimates was computed for each pair of item groups. Scatterplots were also drawn for cross checking.

The Check of the Assumption of Unidimensionality

3.7 The research question is

Are the items of the ITDA Mathematical Test unidimensional?

Bejar's analysis method (1980) was adopted. The 35 items of the Test were sorted into two different content areas[19], namely, Number and Algebra (named as N&A) and Shape, Space & Measures (named as SSM). The items included in each content area are summarized in Table 3:2.

<Table 3:2>

3.8 The RASCAL program was applied to each of the 5 subject groups described in para.3.2. For each subject group, two sets of item difficulty estimates were obtained. One set was obtained by including all items of the ITDA

[19] According to the National Curriculum of U.K. (DfEE & QCA, 1999), "Using & Applying Mathematics", "Number & Algebra", "Shape, Space & Measures" and "Data Handling" are 4 attainment targets. "Using & Applying Mathematics" was ignored as relevant items could be grouped in one of the other content areas. "Data Handling" was not considered as no item was really set on this area in the Test. The only one item in doubt is No.25 which involves arithmetic mean. Since the concept of data handling is not so obvious and this item can be easily solved by simple algebraic method, No.25 is categorized as an item in the "Number & Algebra" content area.

Mathematical Test in the calibration while the other was obtained by estimating only the items within each content area, i.e. N&A and SSM. The two sets of parameters estimated should not differ unless one or both of the content areas is tapping an ability, which is unique to that content area. In applying this method, the following procedures were adopted:

(a) The two sets of the difficulty values in each subject group were correlated (the content-area-based estimates vs. the total-test-based estimates) and plotted, if necessary, to give a preliminary view on the extent to which the two sets are linearly related.

(b) If unidimensionality holds, (i) the plot or the "principal axis" of the item difficulty estimates in (a) should be close to the "theoretical axis" which has slope 1 and intercept 0 (Bejar, 1980, p.284); (ii) the mean distance of items to the principal and theoretical axis should be constant across the two content areas and close to zero. The slope β and intercept α of the principal axis are respectively given by:

$$\beta = \frac{(s_1^2 - s_2^2) \pm \sqrt{(s_1^2 - s_2^2)^2 + 4s_{12}^2}}{2s_{12}} \qquad (3.1)$$

$$\alpha = \overline{B}_1 - \beta \overline{B}_2 \qquad (3.2)$$

where s_1^2 and s_2^2 are the variances of the content-area-based and total-test-based item difficulty estimates respectively, $\overline{B}_1$ and $\overline{B}_2$ are the corresponding mean difficulty estimates, and s_{12} is the covariance between the two sets of estimates. The sign in the numerator is chosen so that β is positive. The proof of equations (3.1) and (3.2) can be found in Appendix 2. The distance of each point of the plot to the principal and theoretical axis can be computed by the equation (A2.3) of Appendix 2:

$$d_i = \left| -B_{2i}\sin\theta + B_{1i}\cos\theta - \alpha\cos\theta \right|$$

$$= \left| \frac{-B_{2i}\tan\theta + B_{1i} - \alpha}{\sec\theta} \right|$$

or $$d_i = \left| \frac{\beta B_{2i} - B_{1i} + \alpha}{\sqrt{\beta^2 + 1}} \right|, \qquad (3.3)$$

where $\beta = \tan\theta$. The corresponding value of d_i for the theoretical axis is found by setting $\beta = 1$ and $\alpha = 0$ in the equation (3.3), i.e.

$$D_i = \left| \frac{B_{2i} - B_{1i}}{\sqrt{2}} \right|. \qquad (3.4)$$

For each content area, the mean distances to the principal axis and to the corresponding theoretical axis were then compared.

3.9 The ITDA mathematics items consist of two forms of multiple-choice items, namely the regular form and the comparative form. Items 1 – 15 are regular items while items 16 – 35 are quantitative comparison items which have the feature of sharing the same response options and instructions:

> Questions 16 – 35 each consist of two quantities, one in Column A and one in Column B. You are to compare the two quantities and on the answer sheet circle the letter
> A if the quantity in Column A is greater;
> B if the quantity in Column B is greater;
> C if the two quantities are equal;
> D if the relationship cannot be determined from the information given.

For example, items 16 and 24 are in the following format:

	Column A	Column B
16.	(-1)(-3)(-5)(-7)	(-1)(-3)(-5)(-7)(-9)
24.	90% of 110% of 8	8

3.10 The feature of quantitative comparison items might lead to multidimensionality due to the particular correct response of the items (Kingston & Dorans, 1985, p.285). Some examinees who did not know the answer might be more likely to choose option D. That means, if the correct answer were A, B or C, some examinees would be less likely to select the correct answer. On the other hand, if D were the correct answer, the same examinees would be more likely to select the correct answer than the model predicted. To investigate the effect of this feature to examinees, Bejar's analysis method was applied to each subject group to get two sets of item difficulty estimates. One set was obtained by including all items in the calibration (the total-test-based estimates) while the other was obtained by estimating only the items of each form (the form-based estimates). The analysis in para.3.8 above was repeated.

The Check of the Assumption of Equal Item Discrimination

3.11 The research question in this section is:

Are the item discrimination levels close to each other?

For each of the 5 subject groups described in para.3.2, the point biserial correlation of each item and the standard deviation of the resulting correlations were computed. The standard deviation should be small if the assumption is held.

The Check of the Assumption of No Guessing

3.12 The research question is:

Are the guessing levels of the items close to zero?

One possible way is to examine the performance of examinees in the low score groups on the four hardest items. For each of these hard items, the proportion of correct responses for each score group in each subject group was plotted against the score group. If the proportions for the low score groups are substantially different from zero, guessing could be assumed to be operating.

The Check of the Assumption of Non-Speededness

3.13 The research question is

Is the ITDA Mathematical Test a speeded test?

The appropriateness of the assumption was checked by investigating the percentage of omits of each item of the Test, the percentage of examinees completing the Test, the percentage of examinees completing 75% of the Test and the percentage of examinees who did not respond to the last 5 items of the Test.

The Comparison between the Classical Test Theory Method and the Rasch Approach

3.14 The CTT model and the Rasch model were compared by addressing the following two major issues as suggested by Fan (1998, p.361):

(a) How comparable are the item and examinee parameter estimates from the CTT framework with those from the Rasch model?

(b) How invariant are the item and examinee parameter estimates of the CTT framework and the Rasch model across participant samples and item groups?

3.15 More specifically, the two issues were expressed as the following four research questions:

(a) How comparable are the CTT-based and Rasch-based item difficulty estimates?

(b) How comparable are the CTT-based and Rasch-based examinee ability estimates?

(c) How invariant are the CTT-based and Rasch-based item difficulty estimates across different participant samples?

(d) How invariant are the CTT-based and Rasch-based examinee ability estimates across different item sets?

Comparability of CTT-based and Rasch-based item difficulty parameters

3.16 Comparability of the two sets of item difficulty parameters was examined by correlating the CTT and Rasch item difficulty estimates obtained from each of the 5 subject groups of examinees. The item difficulty indices p (i.e. the proportion of examinees that responded correctly to the item) were used in CTT and the b-values in the Rasch approach. Since the p-value is an inverse indicator of item difficulty and expresses item difficulty on an ordinal scale (i.e. it indicates the rank order or relative difficulty of items), it is first transformed to an interval scale before conducting statistical analyses. The transformation, which requires the assumption that the underlying trait measured by an item is normally distributed, is achieved by finding the z-score corresponding to the (1 – p)th percentile from the z-distribution. For example, if p = 0.93 (i.e. 93% of the examinees respond to the item correctly), the z-value for such a p-value will be –1.5. This transformation removes the curvilinearity in the relationship between two sets of item p-values (Anastasi & Urbina, 1997, p.174).

Comparability of CTT-based and Rasch-based examinee parameters

3.17 Comparability of the two sets of examinee parameters (raw total score T in the CTT vs. ability θ in the Rasch approach) was assessed by correlating the T and θ estimates obtained from the same subject group of examinees. A high correlation was expected as the raw total score (or the number correct score) has been shown to be a sufficient statistic for the ability θ in Appendix 1. Nevertheless, the CTT-scores T were still correlated with the θ values for justification. The analyses were repeated for all different subject groups described in para.3.2.

The degree of invariance of the CTT-based and Rasch-based item difficulty estimates

3.18 For each of the 25 random samples described in para.3.2, the degree of invariance of the item difficulty parameters was assessed by correlating the estimates obtained within each measurement framework and studying the corresponding scatterplots. As in para.3.16, p-values in the CTT and b-values in the Rasch framework were used. If the correlations and the features of the scatterplots from each measurement framework are close to each other, the CTT and the Rasch modeling should make no significant difference to the invariance of item difficulty estimates.

The degree of invariance of the CTT-based and Rasch-based examinee ability estimates

3.19 For each of the item group described in para.3.6, the examinee ability estimates of the Whole Group obtained within each measurement framework were correlated. Again, if the correlations from the CTT model and the Rasch model and the features of the corresponding scatterplots are close to each other, the two

frameworks should make no significant difference to the invariance of the examinee ability estimates.

Identifying Poor Items using Independent Analyses from the Classical Test Theory Method and the Rasch Approach

3.20 The research question is

> Are the CTT method and the Rasch approach identifying the same "poor" items?

Two indices for each item are required for the item analysis using CTT. They are the p-value and the point biserial correlation coefficient between the scores on the item and the scores on the total test. Kline's suggestions (1990, p.90) were adopted in this study. That is, items to be retained should have a p-value between 0.20 and 0.80 and a point biserial correlation coefficient greater than 0.3. From the point of view of measurement, the purpose of an item is to spread out examinees' scores along a continuum so that examinees could be discriminated. Therefore, a highly discriminating item should have a large spread (or item variance). If an item is answered correctly by most of the examinees (so that it has a high success rate) or by only a few examinees (so that it has a low success rate or a high "failure" rate), its variance will be small and it has made only little discrimination (discriminated only a high or low ability group). Therefore, items should have the highest possible item variance for making a large discrimination to the examinees. Since the variance of an item is given by

$$V = p(1 - p),$$

we have $\frac{dV}{dp} = 1 - 2p$ and $\frac{d^2V}{dp^2} = -2$.

It is clear that p = 0.5 gives a maximum value to the item variance. That is to say, items with equal success and failure rates are maximally discriminating. However, if all the items of a test have p = 0.5, the test could not discriminate very bright examinees and low ability examinees. Very bright examinees would all respond correctly and low ability examinees would all fail. That is why p-values between 0.2 and 0.8 are recommended in a test. Similarly, the total score on a test should measure a broader concept and hence provide more information than any one item. Thus, very high inter-item correlations are not good either. It is because very high correlation must mean that the items are themselves highly correlated. Therefore, they would lead to the production of an exceedingly narrow test and some of the items are virtually identical and hence redundant.

3.21 In Rasch modeling, the fitness of items was checked by examining the chi-square statistic. The "poor items" which were independently identified by the CTT method and the Rasch modeling could then be compared.

CHAPTER 4

RESULTS

Introduction

4.1 The ITDA Mathematical Test in this study requires examinees to complete 35 items in 25 minutes, that is, examinees have on average only 43 seconds for answering each item including reading all the instructions (particularly those for items 16 – 35 which are of quantitative comparison form). Therefore, the time may be tight[20] for most examinees and the assumption of non-speededness may be violated. A preliminary view on the assumption of non-speededness can be obtained by analyzing the percentages of omits of each item. Table 4:1 shows the results. Clearly, more than 10% of examinees did not respond to items 7, 15, 21 – 35 (17 numbers), and more than 20% did not respond to items 25 – 35 (11 numbers). There are two possibilities for the high omitting rate: (a) the items concerned are too difficult so that they were left unanswered; (b) the examinees did not have sufficient time to complete the test. Anyway, the assumption of non-speededness need to be studied in more detail.

<Table 4:1>

4.2 To further study the possible effect of speededness on the estimation of

[20] The time for the multiple-choice paper (of 54 items) in the certificate level mathematics subject of the Hong Kong Certificate of Education Examination (equivalent to GCE "O" Level) is 90 minutes, i.e. 100 seconds for answering each item.

parameters and the results of analysis, the data were re-analysed by deleting from the original test, the last 11 items (i.e. items 25 – 35), which had been omitted by more than 20% of examinees. In the repeated analysis, the same subject groups and samples were used for easy comparison. However, two points should be noted:

(a) Some examinees in the High-ability Group may not be really "high" in score. For example, an examinee with raw score 23 who had all the last 11 items correctly responded to in the original test would only get a score of 12 after taking away the items.

(b) The rationale of the repeated study is based on the assumption that if examinees did not have sufficient time to complete the test, they would omit the items at the end as a block.

For convenience, the original Mathematical Test (of 35 items) is named as Test-35 while that which has deleted the last 11 items, Test-24 (of 24 items). In all the analysis in both tests, the RASCAL program was used and the data were centred on the examinees rather than the items (see footnote 18 on page 38).

The Fit of Rasch Model to the ALIS Data

4.3 The item difficulty and the examinee ability parameters for different participant samples (random, gender and ability samples) were estimated by using the RASCAL program. When the Rasch model fits the test data, the item difficulty parameter estimates will be the same within sampling errors regardless of the samples of examinees chosen from the random samples, gender samples or ability samples. Similarly, the examinee ability parameters will be invariant no matter what item groups are used to estimate the abilities of examinees.

Invariance of item difficulty parameter

4.4 The correlation coefficients between any two sets of item difficulty estimates computed in each sampling plan are listed in Table 4:2 – 4:4.

<Table 4:2 – 4:4>

It can be seen that in both Test-35 and Test-24, all the correlation coefficients are statistically significant at the 0.01 level although the correlation between the high-ability and the low-ability samples are not so "strong" as those among the random samples and those between the male and female samples. In general, the correlation coefficients in Test-24 are slightly smaller than their counterparts in Test-35 except those of r_sample1 and r_sample2, r_sample1 and r_sample3, r_sample1 and r_sample4, h_sample1 and L_sample4, and h_sample5 and L_sample4. Nevertheless, the differences are negligible.

4.5 The correlation between the item difficulty estimates from individual samples were averaged across samples under the same sampling condition to get a general view of the comparability of invariance of item difficulty estimates. In Table 4:5 and all the following tables, an average correlation coefficient was obtained through transforming individual correlation coefficients to Fisher's Zs, averaging the Fisher's Zs and then re-transforming the average value to the Pearson correlation coefficient.

<Table 4:5>

Clearly, the average correlation coefficient decreases across progressively less comparable participant samples. The relationship is weakest in the ability group sampling. The values obtained in Test-24 were slightly less than the corresponding values in Test-35 in all the sampling plans.

4.6 Plots of item difficulty estimates in the two random samples, r_sample3 and r_sample5 which have a correlation coefficient of 0.993 in Test-35 and 0.990 in Test-24 are shown in Figure 4.1(a) & (b). The line "y = x" was added in each of the plots to serve as the baseline. Four other plots from other sampling plans are shown in Figure 4.2 and 4.3. Two compare the estimates from m_sample3 and f_sample2 which have relatively higher correlation coefficients (0.976 in Test-35 and 0.974 in Test-24) and the others compare those from h_sample3 and L_sample2 which have relatively smaller correlation coefficients (0.882 in Test-35 and 0.872 in Test-24). Plots from other pairs of samples can be similarly obtained.

<Figure 4.1 – 4.3>

The plots in Figure 4.1(a) & (b) show a high relationship between the sets of item difficulty estimates. The slight variations in the plots are due to sampling errors. If the feature of item invariance is present, similar results should be obtained from the other plots. Figure 4.2(a) & (b) reveal clearly that the two sets of item difficulty estimates from m_sample3 and f_sample2 differ a bit in both tests and the scatter in each is a little bit below the baseline (with Test-24 slightly better). Figure 4.3(a) & (b), on the other hand, differ substantially from the plots shown in Figure 4.1 – 4.2 (a) & (b). The points are more scattered and well below the baseline. Although the large distance from the

line of the low ability group is partly the result of the data being centred on the examinees, Figure 4.1 to 4.3 suggest that the test items function a bit differently in the two ability samples in both Test-35 and Test-24. There are three possible explanations for the difference:

(a) The item parameter estimation is not done very well when extreme groups are used as some effects like speededness, carelessness and nervousness on the test performance of examinees, particularly the low ability examinees were not taken into account.

(b) Other important item statistics like item discrimination index and guessing factor have been ignored. That is to say, the problem may be due to model-data misfit.

(c) Some other assumptions of the Rasch model are violated.

No matter which one, the feature of item parameter invariance is not obtained.

4.7 For cross checking, the chi-square statistics for each item in all the 25 samples are computed for both Test-35 and Test-24. The fitness of each item in each sample was tested at the 0.01 level. Table 4:6 summarizes all the unfitted items in all the samples. For simplicity, only items which are found unfitted in at least 3 samples from each sampling group are considered as "really" unfitted items. These "really" unfitted items are also listed in Table 4:6 for reference.

<Table 4:6>

Table 4:6 shows that in both Test-35 and Test-24, the random samples and the male samples give nearly the same unfitted items but the others give substantially different results. In particular, a large number of items in the

samples of the two ability groups fit the Rasch model. There are only two "really" unfitted items in the High-ability samples in both Test-35 and Test-24 while in the Low-ability samples, there are two "really" unfitted items in Test-35 but none in Test-24. However, the apparent fitness seems to contradict the plots obtained in para.4.6. Therefore, checking of the assumptions should be studied and this will be done in the next sections.

4.8 Regardless of the values of the chi-square statistic, the differences in item difficulty estimates of the Male Group, Female Group, High-ability Group and Low-ability Group relative to the Whole Group which was taken as the control or reference group, were compared by computing the "effect sizes" of the items. The means and standard deviations of the item difficulty estimates of all the 5 subject groups as well as the effect sizes are tabulated in Table 4:7.

<Table 4:7>

It is seen that the means of item difficulty are different for different groups and the differences of those of the two ability groups are more substantial (-1.100 in the High-ability Group and 2.924 in the Low-ability Group vs. 0.584 in the Whole Group). Similarly, the item difficulty estimates of the two ability groups (which have relatively higher standard deviations of 2.304 for the High-ability Group and 1.872 for the Low-ability Group) are more scattered when compared with those in the Whole Group and the gender groups.

<Figure 4.4>

The scatterplots of the effect sizes are plotted in Figure 4.4. The figure reveals that the effect sizes of the items of the two ability groups are substantially different from those of the two gender groups. Most of the effect sizes of the items of the High-ability Group are negative showing that the item difficulty estimates are generally smaller than those estimated from the Whole Group. That is to say, most of the items are easy to this group of examinees. On the other hand, the estimated item difficulty parameters of the Low-ability Group are substantially higher than those estimated from the Whole Group or most of the items are hard to the low-ability examinees. The effect sizes of the two gender groups are relatively smaller (with those from the Female Group a little bit above the horizontal axis and those from the Male Group a little bit below) showing that the "effect" of the gender is not so substantial.

4.9 For further study, the items are ordered in terms of the item difficulty parameters estimated from various subject groups and tabulated in Table 4:8. In Test-35, the first 4 simplest items are in the order of 2, 1, 8, 5 in all groups except the Female one in which the order is 2, 8, 5, 1. The hardest 3 items are 33, 34 and 35 in the Whole Group, Male Group and High-ability Group although they are in different order of difficulty. Different combinations were identified in the other two groups. The items are 13, 34 and 35 in the Female Group while in the Low-ability Group, the order is 34, 30 and 35. Apart from that, the orders of other items in all the subject groups are quite different particularly those in the two ability groups (the Whole Group and Male Group being a bit better).

<Table 4:8>

Similar results are obtained from Test-24. The first 4 simplest items have the same pattern as those in Test-35. However, the hardest items identified are 14, 15 and 13 and they are in the same order in all groups except the High-ability Group in which the order is 15, 14, 13. Again, the orders of other items are different in the Female Group and the two ability groups. On the other hand, the orders in the Whole Group and Male Group are more or less the same. If the orders of the items in Test-35 and Test-24 are compared in each group, it is not hard to see that the order of the first 24 items of Test-35 remains the same as those in Test-24 no matter what group is considered. Table 4:8 reflects that not only the invariance of the item difficulty estimates is not preserved, but also that of the order of the items.

Invariance of ability parameter

4.10 To investigate the invariance of ability parameters across different item groups (i.e Equivalent-halves Groups, Content Groups and Difficulty Groups), the item difficulty parameters of the items in both Test-35 and Test-24 were estimated by applying the RASCAL program to the examinees in the Whole Group. Table 4:9 shows the item difficulty estimates of the items of Test-35 and Test-24 and also the estimates sorted in item difficulty order for completeness. Figure 4.5 (a) & (b) give the corresponding graphs showing the item difficulties arranged by the order of the items. As stated in para.4.9, in terms of difficulty, the order of the first 24 items remains the same in both Test-35 and Test-24.

<Table 4:9>

<Figure 4.5>

Items with negative difficulty estimates were considered as easy items while those with positive values hard items. The results are tabulated in Table 4:10 and they reflect that both Test-35 and Test-24 identify the same easy items.

<Table 4:10 – 4:11>

4.11 The correlation coefficients of examinee ability were estimated from the Equivalent-halves Groups, Content Groups and Difficulty Groups using the RASCAL program. The correlation coefficients of the estimates from each pair of item groups was then computed. The results are listed in Table 4:11. The correlation coefficients are fairly constant across the Equivalent-halves Groups and Content Groups (0.580 & 0.527 respectively in Test-35 and 0.342 & 0.285 respectively in Test-24) but those obtained from the Difficulty Groups were substantially smaller (0.013 in Test-35 and 0.023 in Test-24). Since the degree of heterogeneity of samples will influence the values of correlation coefficients (Hopkins et al, 1996, p.112), the low values obtained in the Difficulty Groups may be due to the "less heterogeneity" of the abilities of examinee estimated by the easy items (and also by the hard items). Nevertheless, the low correlation coefficients, to an extent, indicate that the examinee ability estimates are not invariant across different sets of items. The scatterplots in Figure 4.6 – 4.8 provide visual comparisons between ability estimates obtained with different item groups.

<Figure 4.6 – 4.8>

From the plots, it is found that most points are clustered about the baseline except the extreme values, i.e. the examinees with very low abilities (with abilities equal –9). Obviously, these extreme values, being isolated from the clustered points, may give biased results as the influence of speededness and guessing is believed to be most significant for examinees with low abilities. In addition, there are more extreme values in the scatterplots based on items of different difficulties (i.e. Figure 4.8) (1301 & 220 numbers out of 26,964 ability estimates in Test-35; and 1301 & 653 numbers in Test-24) than the others (i.e. Figure 4.6 & 4.7). For further study, the correlation coefficients were re-calculated by deleting the extreme values and are also tabulated in Table 4:11. The results reflect that the correlation coefficients are increased substantially, particularly in the Difficulty Groups. The values in Test-35 and Test-24 are increased from 0.013 and 0.023 to 0.477 and 0.421 respectively. The correlation coefficients were relatively more constant across the item groups in both Test-35 and Test-24 after deleting the extreme values.

The Check of the Assumption of Unidimensionality

4.12 Bejar's analysis method was adopted to check the assumption of unidimensionality. For simplicity, the 5 groups of data (i.e. the Whole Group, Male Group, Female Group, High-ability Group and Low-ability Group) were tested instead of the 25 samples. The RASCAL program was applied to the Whole Group twice to obtain two sets of item difficulty estimates. One set (total-test-based) was obtained by including all 35 items of the test concerned while the other (content-area-based) was estimated by only items within each content area, i.e. N&A and SSM. The two sets of estimates hence obtained

were then correlated. The procedures were repeated for the other 4 groups. The correlation coefficient of the two sets of estimated item difficulty parameters of each individual group in each test is tabulated in Table 4:12.

<Table 4:12>

It is seen that in Test-35, the correlation between the two sets of estimates is nearly "perfect" in all the groups except the Low-ability Group which has a value of 0.982. Although the correlation coefficient for this group is still high, its anomaly can be reflected from the plot of the two sets of item difficulty estimates as shown in Figure 4.13(a). The corresponding plots for the other groups are also shown in Figure 4.9(a) – 4.12(a) for comparison. Clearly, the two sets of estimated parameters for the Low-ability Group tended to form a pattern of "fan" and deviated severely for the items with high difficulty level. Test-24 shows improvement. There is no substantial difference for the Whole Group, the two gender groups and the High-ability Group, while the deviation of the two sets of item difficulty estimates in the Low-ability Group is much smaller when compared with that in Test-35.

<Figure 4.9 – 4.13>

The anomaly of the low-ability group was further studied by considering the slope, β and intercept, α of the principal axis of the estimated item difficulty parameters. If unidimensionality holds, β will be close to 1 while α to 0. Table 4:13 shows the values of β and α for various groups (calculated from equations (3.1) and (3.2)). It is seen that except the High-ability Group and

Low-ability Group (in particular, the Low-ability Group in Test-35), the principal axes of all other groups are close to the theoretical axis. This result also agrees quite well with the plots in Figure 4.9 – 4.13.

<Table 4:13>

4.13 The mean distances of items to the principal and theoretical axis were computed for each content area and subject group and are tabulated in Table 4:14. Figure 4.14(a) & (b) plot the same information.

<Table 4:14>

<Figure 4.14>

The principal axis relating to the content-area-based and total-test-based estimates should be close to the theoretical axis and the average distance to the theoretical axis should be small and constant across the two content areas. It is not hard to envisage that there would be great differences in distance across content areas even the two axes coincide. Therefore, both requirements must be met for unidimensionality. The results indicate that the responses to Test-35 are accounted for quite well by a single dimension in the Whole Group and the two gender groups. For the two ability groups, particularly the Low-ability Group, the principal and theoretical axis do not coincide as well as they do in other subject groups. Moreover, there are great differences in the mean distances whereas such distances are relatively smaller and constant in other groups. This indicates a severity of departure from unidimensionality in the

two extreme ability groups. The great difference between the mean distances from the theoretical and principal axes in the two ability groups disappeared in Test-24, but their mean distances are still relatively larger than those obtained in other groups.

4.14 Bejar's procedures were re-applied to the 5 subject groups to study the effect of different forms of multiple-choice items (regular and quantitative comparison forms) on unidimensionality. The corresponding tables and plots are shown in Table 4:15 – 4:17 and Figure 4.15 – 4.20 respectively.

<Table 4:15 – 4:17>

<Figure 4.15 – 4.20>

The correlation coefficients between the form-based estimates and the corresponding total-test-based estimates are high (greater than 0.9) and those in Test-24 are relatively higher than those in Test-35 except the High-ability Group. Nevertheless, as in the case of different content categories, the points substantially deviate from the baseline in the Low-ability Group and at the "lower end" of the High-ability Group. These results were justified by considering the slopes and intercepts of the theoretical and principal axes and the mean distances to the two axes for each item form. From Figure 4.20, it is not hard to see that the result has been improved after deleting the last 11 items although the mean distances in the two ability groups are still large.

The Check of the Assumption of Equal Item Discrimination

4.15 The discrimination of each item is measured by the point biserial correlation. For each of the 5 subject groups, the point biserial correlations of all the 35 items in Test-35 and the 24 items in Test-24 were respectively computed and tabulated in Table 4:18.

<Table 4:18>

The tabled entries reflect that the means and standard deviations are fairly constant across the Whole Group, Male Group and Female Group (ranging from 0.313 to 0.338 in Test-35 and 0.356 to 0.379 in Test-24) in both tests, while those of the ability groups deviate from each other and also from the first three groups. This result indicates that the item discrimination depends on the groups of examinees and is by no means identical. The High-ability Group and Low-ability Group have relatively smaller standard deviations (0.081 and 0.058 in Test-35; 0.078 and 0.063 in Test-24). Therefore, the variation in the item discrimination indices is smallest. This may be explained by the fact that the items, in general, are relatively too easy to the high-ability examinees and too difficult to the low-ability ones so that all items look more or less "the same" to the examinees in each ability group in terms of discrimination. The variations in other subject groups are relatively more substantial.

4.16 Glancing through the point biserial correlations of the last 5 items in Test-35, it is not hard to see that with a few exceptions (say item 2 which is the easiest item to all subject groups as seen in Table 4:8 and hence has the lowest item

discrimination as expected), the values of their correlations are generally smaller than those of other items in the Whole Group and the two gender groups. The low item discriminations of these items may attribute to the speededness of the test. As most students (see para.4.20 below) could not get sufficient time to achieve the last few items, examinees either left them unanswered or guessed resulting in low item discriminations and hence low point biserial correlations.

The Check of the Assumption of No Guessing

4.17 The guessing factor is checked by considering the performance of examinees on the hardest few items in the low score groups. The observed proportions of correct responses to the items in both Test-35 and Test-24 for each score group were computed for the Whole Group, Male Group, Female Group, High-ability Group and Low-ability Group. Although the lowest score in the High-ability Group is 18 in Test-35 by definition, it is included here for completeness. For comparison purpose, the expected ICCs were also included. The 6 hardest items and their item difficulty parameters estimated for each subject group in each of Test-35 and Test-24 are presented in Table 4:19.

<Table 4:19>

For comparison purposes, the same four items were studied in all the subject groups in each test (i.e. items 13, 33, 34 and 35 in Test-35 and items 7, 13, 14 and 15 in Test-24). The plots of proportions of correct responses to these "hard" items are shown in Figure 4.21 – 4.24 (Test-35) and Figure 4.25 – 4.28 (Test-24).

<Figure 4.21 – 4.28>

4.18 It is seen that the observed proportions correct are higher than the expected values for score groups roughly lower than 20-25 in Test-35 and 15-20 in Test-24. In addition, for score groups, say from 10 to 20 in Test-35 or from 5 to 10 in Test-24, the observed proportions correct of most items remain fairly constant (for example, items 33, 34 and 35 in Test-35 and 13, 14 and 15 in Test-24 of the Whole Group, Male Group and Female Group). In particular, the observed ICC for items 34 in Test-35 and those for items 7 and 14 in Test-24 reflect a substantial variation from their expected counterparts. The observed ICCs at low score groups are higher than the expected ones but the reverse is observed at high score groups. This reflects that there might be some ambiguous wordings or misconcepts which mostly affected examinees in high score groups. Generally, the assumption of zero guessing factor may be incorrect. For the Low-ability Group, most of the proportions are less than 0.10 in Test-35 and between 0.10 and 0.20 in Test-24.

The Check of the Assumption of Non-Speededness

4.19 Response speed is as much an indicator of ability as is the correctness of responses to items. It is a different dimension which affects performance as examinees who can not only do mathematics, but also do it quickly will tend to get higher grades than those who need more time. Therefore, speed and ability are sometimes confounded. Nevertheless, the effect of speededness in this study was investigated. The preliminary analysis in Table 4:1 has shown that the ITDA Mathematical Test might be a speeded test to the group of examinees

concerned. To seek for further justification, the percentages of examinees completing Test-35 and completing 75% of the test were computed and are presented in Table 4:20.

<Table 4:20>

The results indicate that even in the High-ability Group only half of the examinees can complete the whole test and 91.5% three quarters of the Test. The corresponding values for the examinees in the Low-ability Group are even much lower (29.0% and 53.2% respectively).

4.20 The assumption of non-speededness was further studied by considering the percentages of examinees who did not respond to the last 5 items, i.e. items 31 to 35. The values are tabulated in Table 4:21. It is seen that over 20% of examinees in all subject groups (except the High-ability Group which has a relatively smaller percentage of 16.4) have omitted all the last 5 items. In particular, the Low-ability Group has a very high value of 46.7%. Moreover, the percentages increase from item 31 to item 35 for all groups (except item 34 and 35 of the Low-ability Group which have roughly the same percentage). Therefore, the examinees had insufficient time to complete the test and a large percentage of them omitted the last few items. According to Slinde & Linn (1979, p.441), a non-speeded test is considered to be one in which the proportion omitting the last few items is about 0.10 or less. Therefore, all the data obtained have provided an indication of the speededness of the ITDA Mathematical Test.

<Table 4:21>

The Comparison between the Classical Test Theory Method and the Rasch Approach

Comparability of CTT-based and Rasch-based item difficulty parameters

4.21 Table 4:22 presents the results, associated with the first research question of para.3.15, that is, "How comparable are the CTT-based and Rasch-based item difficulty estimates?". For each of the subject group, the normalized or transformed CTT p-values and the Rasch-based item difficulty estimates were correlated.

<Table 4:22 – 4:23>

As the results indicate, the relationship between CTT- and Rasch-based item difficulty estimates is almost perfect for all subject groups in both Test-35 and Test-24 because the correlation coefficients are either 0.999 or 1.000. For cross-checking, the orders of the items based on the CTT and Rasch item difficulty estimates are tabulated in Table 4:23. It can be seen that in Test-35, except items 16 & 19 in the Female Group and items 4 & 11 in the Low-ability Group, the two measurement frameworks created the same item order. In Test-24, similar result was obtained. It appears that both the CTT and Rasch model provide almost the same information with regard to item difficulty but the latter has a considerable model complexity. Unless the Rasch model estimates could show superior performance in terms of invariance across different samples over that of CTT item difficulty indices, the results obtained here might not indicate any empirical advantage over the simpler CTT framework.

Comparability of CTT-based and Rasch-based examinee parameters

4.22 For the second research question of para.3.15, that is, "How comparable are the CTT-based and Rasch-based examinee ability estimates?", both CTT- and Rasch-based ability estimates (i.e. the raw total scores and the θ estimates) were obtained and correlated for each of the subject groups. Table 4:24 presents the results for the groups. The results show that the CTT- and Rasch-based ability estimates correlate highly with each other for all subject groups in Test-35. All the correlation coefficients have values of at least 0.95 except the High-ability Group which has a lower value of 0.889. The correlation coefficients of the subject groups (except the Low-ability Group) are smaller in Test-24 than in Test-35. In particular, the High-ability Group has an exceptionally lower correlation coefficient of 0.714. Nevertheless, the correlations are still high. These high correlations indicate that CTT- and Rasch-based ability estimates are comparable with each other. That is to say, regardless of which measurement framework is used, very similar conclusions will be drawn. This has justified our expectation as the raw total score is a sufficient statistic of the θ estimate in Rasch model.

<Table 4:24>

The degree of invariance of the CTT-based and Rasch-based item difficulty estimates

4.23 The third research question in para.3.15 is "How invariant are the CTT-based and Rasch-based item difficulty estimates across different participant samples?". It is a very crucial question as the assumption of item parameter invariance across different participating samples will justify the use of Rasch model in measurement. Table 4:25 presents the results for this research question. Notice that the correlation coefficients in this table are averages of the

correlation coefficients between item difficulty estimates from any two different samples (e.g. r_sample1 and r_sample2, h_sample1 and L_sample2, etc.) derived from the same measurement framework.

<Table 4:25>

The tabled entries indicate that the average between-sample correlation coefficients of item difficulty estimates are very high and are comparable between CTT and Rasch model (those in Test-24 are slightly smaller than their counterparts in Test-35). The standard deviations are smallest in the ability group sampling in both Test-35 and Test-24 within each measurement framework (Test-35: 0.034 in CTT and 0.039 in Rasch; Test-24: 0.062 in CTT and 0.041 in Rasch). On the other hand, the standard deviations in the gender group sampling in Test-24 are substantially different in the two analyses (being 0.177 in CTT and 0.072 in Rasch as shown in Table 4:25). The scatterplots of the CTT transformed p-values of the same samples adopted by Rasch model in para.4.6 are shown in Figure 4.29 – 4.31. By comparing these plots with those in Figure 4.1 – 4.3, it is found that the plots exhibit very similar features as the Rasch item difficulty estimates, i.e. a high relationship for the random samples and a slightly weaker relationship for the gender samples, but the points in the plots of ability samples fell substantially below the baseline.

<Figure 4.29 – 4.31>

The degree of invariance of the CTT-based and Rasch-based examinee ability estimates

4.24 The last research question in para.3.15 is "How invariant are the CTT-based and

Rasch-based examinee ability estimates across different item sets?". This question is as crucial as the third research question. Table 4:26 presents all the correlation coefficients across the item groups in both Test-35 and Test-24 within each measurement framework. The results reflect that the correlation coefficients of examinee ability estimates from CTT are relatively more constant across all the item groups than those estimated from the Rasch model, particularly those in the Difficulty Groups (being 0.013 in Test-35 and 0.023 in Test-24).

<Table 4:26>

The corresponding scatterplots of the standard scores (derived from the raw total scores) of the examinees of the Whole Group in different item groups are shown in Figure 4.32 – 4.34. Comparing the plots with those in Figure 4.6 – 4.8, it is not hard to see that the isolated extreme values do not exist, therefore, resulting higher values of correlation coefficients.

<Figure 4.32 – 4.34>

For comparison purposes, the isolated extreme values in the Rasch estimates were excluded with correlation coefficients re-calculated. The results (Table 4:26) reflect that the correlation coefficients within the CTT framework decrease slightly while those from the Rasch framework increase significantly. As a result, the ability parameters estimated from the two frameworks are highly comparable. These empirical observations from para.4.23 and this paragraph are quite interesting as invariance is a strong argument in favor of the Rasch model.

Identifying Poor Items using Independent Analyses from the Classical Test Theory Method and the Rasch Approach

4.25 The CTT p-values and the point biserial correlation coefficients of the items were computed. Using Kline's criteria (1990, p.90), items with p-values between 0.20 and 0.80 and point biserial correlation coefficients greater than 0.3 are "good" items. On the other hand, the poor items were also identified by using chi-square statistics in the Rasch model. The comparability of the numbers of unfitted items from the two measurement frameworks and the number (and percentage) of common unfitted items are tabulated in Table 4:27(a) & (b). The tabled entries show that in all the samples of Test-35, the CTT gives more unfitted items (ranging from 15 to all items) than the Rasch modeling (ranging from 1 to 15 items). In particular, nearly all the items in the high-ability samples and all the items in the low-ability samples are unfitted when assessed by the CTT. In Test-24, the numbers of unfitted items identified by both frameworks are closer to one another than those in Test-35 in the random and gender samples (4 to 8 unfitted items in CTT and 3 to 8 unfitted items in Rasch). However, the great discrepancies in the ability samples (18 to 23 unfitted items in CTT and 0 to 3 unfitted items in Rasch) remain in Test-24. The percentages of commonly unfitted items identified by both frameworks are low in both tests particularly in the low-ability samples (8.6% to 20.0% in Test-35 and 0% to 16.7% in Test-24). Clearly, the two measurement frameworks have identified different "poor" items in this study and the CTT seems to be more sensitive to the lack of fit than the Rasch Model particularly in the two extreme ability groups.

<Table 4:27>

CHAPTER 5

SUMMARY, DISCUSSION AND RECOMMENDATIONS

Summary

5.1 In the present study, the Rasch model which claims person-free item calibration and item-free person measurement was used to analyze the responses of 26,964 examinees to the International Test of Developed Ability (ITDA) Mathematical Test, a test of 35 multiple-choice items. As over 20% of examinees did not respond to the last 11 items, the analysis was repeated by deleting from the full test of 35 items (here called Test-35) the last 11 items to produce Test-24 for further studying the effect of speededness.

5.2 There were two purposes of the study: (a) to test the fit between the Rasch model and the data from the ITDA Mathematical Test, the scores of which provides an alternative baseline for comparing the progress of students in the Advanced-level Information System (ALIS) project in U.K.; and (b) to compare the two popular measurement frameworks, namely the Classical Test Theory (CTT) method and the Rasch approach. To investigate these questions, 5 subject groups, namely the Whole Group (of all 26,964 examinees), Male Group (of 12,567 examinees), Female Group (of 14,397 examinees), High-ability Group (of 6,601 examinees) and Low-ability Group (6,821 examinees), and 25 random samples of size 1000 (5 from each group) were generated. The test items were also regrouped in three different ways to generate 3 pairs of item groups, namely the Equivalent-halve Groups (odd vs. even numbered items),

Content Groups ("Number & Algebra" vs. "Shape, Space & Measures" content areas) and Difficulty Groups (easy vs. hard items). The RASCAL program is adopted to analyze the responses of examinees in various groups/samples to get sets of unconditional maximum likelihood estimates. The major findings are summarized as follows:

(a) The invariance of item parameters was not supported. The plots of item difficulty estimates from the ability samples showed that the points are more scattered and well below the baseline when compared with the plots of random and gender samples. Although the large distance of the scatter from the line is partly the result of data being centred on the examinees, the plots showed that the feature of invariance was not preserved in the samples of extreme abilities. More than that, the difficulty order of the items was found different in different subject groups. Even if scores of the 11 items with high percentages of omits (over 20%) were excluded, there was no significant improvement in the invariance of the item parameters in the ability samples. On the other hand, the difficulty order of the first 24 items in both Test-35 and Test-24 are the same in each subject group.

(b) A fairly good number of items in both Test-35 and Test-24 in the random samples (71.4% in Test-35 and 75% in Test-24) and gender samples (ranging from 71.4% to 87.5%) were found to fit the Rasch model. The fit looked extremely good in the ability samples, but over-fitness may indicate a constraint on the responses (Meijer & Sijtsma, 2001, p.823). In this study, the percentages of omits were high for most items. Therefore, the accuracy of the results in the low-ability samples is imputable to considering all omitted and not-reached items as "wrong"

items. Regardless of the statistical significance of the item difficulty estimates, a study of the effect sizes of individual items of the gender and ability groups (taking the Whole Group as the control or reference group) showed that the "effect" of examinee ability on the item difficulty estimates is substantially larger than that of examinee gender.

(c) The invariance of examinee parameters was not supported by the data. The correlation coefficient between the examinee ability estimates was significantly smaller in the Difficulty Groups than those in the other two pairs of item groups. On the other hand, after taking away the extreme values of those examinees with abilities equal to –9 (ranging from 109 to 1,301 numbers out of 26,964 examinees) who might have omitted a large percentage of items, the correlation coefficients became more constant across the 3 pairs of item groups.

(d) The assumption of unidimensionality, checked by Bejar's analysis method, was not supported in the two ability groups, particularly the Low-ability Group. The two sets of item difficulty estimated in the Low-ability Group, one from the total test while the other from either items within each of the two content areas identified (i.e. "Number & Algebra" and "Shape, Space & Measures" content areas) or items within the two item forms (i.e. regular and quantitative comparison forms), were found to deviate substantially from the baseline.

(e) The assumption of equal discrimination level was not supported in the two ability groups as the means of the point biserial correlation of the items deviated from each other and also from the other three subject groups.

(f) The assumption of zero guessing factor was in doubt as the observed

proportions of correct responses in the low score groups were non-zero for all the subject groups.

(g) The ITDA Mathematical Test might be a speeded test to the group of 26,964 examinees under study as only half of the examinees could complete the whole test and 91.5% complete three quarters of the test even in the High-ability Group.

(h) The item difficulty parameters estimated from the CTT were highly comparable with those estimated from the Rasch model as the two sets of item parameters estimated from the two measurement frameworks were highly correlated in each subject group.

(i) The examinee abilities estimated from the CTT and those from the Rasch model were very comparable as the two sets of ability parameters estimated from the two measurement frameworks were highly correlated in each subject group.

(j) The CTT transformed p-values and the Rasch-based item difficulty estimates exhibited very similar features regarding invariance across samples when scatterplots of item difficulty estimates of the same samples were examined.

(k) The examinee parameters estimated from the CTT appeared to be more invariant across each of the 3 pairs of item groups than those obtained from the Rasch model. However, when the extreme values (abilities equal to –9) were deleted from the analysis, the invariance of the Rasch-based ability estimates across different pairs of item groups was greatly improved and both frameworks exhibited a high degree of invariance.

(l) The CTT method and the Rasch approach identified different "poor"

items. In addition, more items were "found" unfit by the CTT method than the Rasch approach in both Test-35 and Test-24, showing that the former is more sensitive to the lack of fit than the latter.

Excluding the scores of the last 11 items in Test-35 to reduce the effect of speededness did show slight improvement in some aspects like the number of unfitted items (e.g. 28.6% of items were found unfitted in Test-35 vs. 25.0% in Test-24 in the random sampling plan) and unidimensionality in the Low-ability Group. However, the findings in other aspects like comparability of CTT- and Rasch-based item and examinee parameters remained more or less the same.

Discussion

5.3 In this study, the unconditional maximum likelihood procedure was used with the Rasch model. It provided point estimates for all item and examinee parameters. If the Rasch model holds true for the set of ITDA data, the item parameters are invariant or stable when estimated across different groups of examinees and the examinee parameters are invariant or stable when estimated across different item groups. This is the basic rationale for the statistical methods of investigating fit to the Rasch model (Gustafsson, 1980, p.209; Slinde & Linn, 1979, p.441). The analysis of this study showed that the feature of invariance could not be preserved in the groups of extreme abilities. The possible reasons are:

(a) *Biased Estimation Procedures* – The estimation procedures of item and ability parameters were biased.

(b) *Violation of Model Assumptions* – Some of the assumptions about the nature of the data were not satisfied.

(c) *Inappropriate Model* – The Rasch model is not appropriate for calibrating the ITDA mathematics items and measuring the abilities of the examinees.

Biased Estimation Procedures

5.4 The following two procedures adopted in the analysis might have introduced biased estimates:

(i) Items with totally correct or totally incorrect responses and examinees with perfect or zero scores were excluded from the analysis.

(ii) Omitted and not-reached responses were scored as zero in the RASCAL program.

Since the sample size of the data set (26,964 numbers) was relatively large, the exclusion of items with totally correct or totally incorrect responses and examinees with perfect and zero scores should have been reduced to a minimum. On the other hand, since the percentages of omits (including the not-reached items) were large in the data set, it had definitely produced biased estimates. Lord (1980, pp.226-229) suggested that the not-reached items could be ignored (assuming that all not-reached responses fall in a block at the end of the test) and presumably supply random responses in place of omits. However, there are two deficiencies in the method. First, some examinees did not respond to the items in serial order so that mistakes would be made in assuming that not-reached items only occur at the end of the test. Second, presumably supplying random responses in the place of omits would introduce a guessing factor to the responses. This contradicts to the assumption of zero guessing factor of the Rasch model.

5.5 In this study, the unconditional maximum likelihood procedure was adopted to analyze the responses of examinees. However, conditional maximum likelihood (CML), marginal maximum likelihood (MML) and Bayesian estimation procedures could also be used. In the CML estimation, the likelihood function is expressed in terms of the raw total scores and not the unknown abilities of examinees. It is possible as the raw total score is a sufficient statistic for the ability level in the Rasch model (Hambleton & Swaminathan, 1985, pp.138-139; Embretson & Reise, 2000, pp.214-215). CML estimators are consistent and normally distributed (cited in Embretson & Reise, 2000, p.217). However, the CML estimation does not have no drawback. For example, the precision in computing the elementary symmetric functions is difficult for long tests. Also, like the unconditional maximum likelihood procedure, no estimates are available for items or examinees with zero or perfect scores. In the MML procedure, the observed data (the responses) are regarded as a random sample of the population. Unknown ability level is assumed to be a continuous variable with a specific shape distribution in the population, typically a normal distribution, and is handled by expressing the response pattern probabilities as expectations from the population distribution. Nevertheless, the distribution need not be known in advance. If sufficient sample size is available, the ability distribution may be estimated from the data. One of the advantages of MML is the availability of estimates of zero and perfect scores. Its main disadvantage is that a distribution must be assumed for the ability level, thus making the parameter estimates contingent on the appropriateness of the assumed distribution. Like the CML estimation, it does not produce person parameter estimates directly. In a Bayesian method, prior information on the distribution of ability parameters is available based on either

theoretical or empirical considerations. Since the same ITDA Mathematical Test has been used in the ALIS project since 1988, a good picture of the frequency distribution of ability of the group of examinees to be tested can be formed. Such prior information can be used to improve parameter estimation. In sum, further research could be done with these three different estimation procedures.

Violation of Model Assumptions

5.6 The analysis showed that the basic assumptions (i.e. undimensionality, equal discrimination level, zero guessing factor and non-speededness) were not satisfied particularly in the two ability groups (i.e. High-ability Group and Low-ability Group). Among all, the violation of the assumption of non-speededness was more serious although speed and ability are sometimes confounded. The testing time was only 25 minutes which was inadequate for most examinees. Examinees of the same ability scored differently as slow examinees did not even get a chance to attempt the items late in the test. The items concerned therefore might appear to have higher discrimination (Mead, 1976, p.9; Anastasi & Urbina, 1997, p.193). However, with the influence of guessing, the result will be very complicated.

5.7 To eliminate the influence of speed, one possible way is to limit the analysis of each item to those examinees who have reached the item. However, since the number of subjects attempting the later items will shrink rapidly and hence render the results quite unreliable. This procedure is not completely satisfactory unless the number of examinees failing to reach the item is small (which is clearly not true in this study as over 20% of examinees did not respond

to the last 11 items). Moreover, the faster performers tend to be the more proficient and hence the later items would be analyzed on a superior sample of examinees. This would lower the apparent difficulty level of the later items as the percentage correct would be greater in this superior group than in the entire group. This procedure has also some effects on the item discrimination indices. Since slow examinees have a greater tendency to guess as they make effort to try all items within the time allowed, in the absence of these examinees, the sample on which the later items are analyzed will cover a relatively narrow range of ability. The discrimination indices of the later items will therefore tend to be lower than they would be if computed on the entire sample.

5.8 The violation of the assumption of non-speededness is not easy to deal with. Several models have been proposed to incorporate speed into the estimation of ability level (e.g. Rasch, 1980, pp.34-49; Meredith, 1970, pp.49-82; Roskam, 1997, pp.187-208; Verhelst et. al., 1997, pp.169-185). In these models, the probability or probability density of the response time to an item is essentially considered as a function of the item difficulty and the examinee's mental ability (which may be regarded as a combination of power and speed) assuming an exponential response time distribution (Roskam, 1997, p.187). Further research on the fitness of data could be done with this speed and time-limited test.

5.9 Since the ITDA Mathematical Test is a speeded test, it is clearly not unidimensional as the examinee's speed is also an essential "trait" affecting his/her performance. The results of Bejar's analysis method indicated that the assumption of unidimensionality held in all subject groups except that of the

Low-ability. This provided evidence to the anomaly of the plot of content-area-based difficulty estimates vs. total-test-based estimates of the Low-ability Group which was mostly affected by speededness.

5.10 To low-ability examinees, the item format should be simple and straightforward to avoid any discomfort introduced because they are mostly affected by peculiarities. In particular, when testing time was limited, examinees in the Low-ability Group might have suffered from anxiety, carelessness and misunderstanding resulting from the change of item form. This might explain why the effect of different item forms (namely, the regular form and the comparative form[21]) was significant in the Low-ability Group.

5.11 The speededness of the ITDA Mathematical Test naturally encouraged examinees to guess, particularly those of lower abilities. The Item Characteristic Curves (ICCs) of some items (say items 13, 33 – 35 of Test 35 and items 7, 13 – 15 of Test 24) revealed that the guessing factor lies roughly between 0.1 to 0.2 for score groups of 5 – 20. For the lowest score groups, say 2 and 3, the examinees might not even care and they just left the items unanswered. This might explain why the observed and expected ICCs nearly coincided in these low score groups for the Whole Group, Male Group, Female Group and Low-ability Group.

5.12 In particular, the observed ICCs of item 34 in Test-35 and items 7 and 14 in Test-24 deviated most from the expected ICCs. Examinees of low ability had

[21] Items of comparative form have the features of comparing the quantities of two expressions and sharing the same response options and instructions.

higher observed proportions correct while those of high ability had lower observed proportions correct than the expected ones. These peculiarities need further study of the items. Item 7 is a problem about simultaneous equations in two unknowns. Students are requested to find the value of r given that $r = \frac{3}{4}s + 6$ and $\frac{3}{2}s + 12 = 5$. The fractions $\frac{3}{4}$ and $\frac{3}{2}$ might have confused some examinees. High-ability examinees might make careless mistakes while low-ability examinees might just guess or leave the item un-attempted (12.8% of examinees did not respond to this item). Item 14 is a simple trigonometric problem. Students are asked to calculate the length of AC in cm (see Figure 5.1). The answer is simply $2 \times 8\sin 60°$, but some examinees might have approached the problem by first finding the lengths of AD and BD from ΔABD and then get the length of AC by applying Pythagoras' Theorem to ΔACD. In either method, confusion between sine and cosine functions might introduce mistakes to high-ability examinees. (In fact, this is also a common mistake normally made by students in Trigonometry.) Again, item 34 is also a simple simultaneous linear equations problem but it is of comparative form. Students are requested to compare the values of $x + y$ and 1 given that $3x - 4y = 2$ and $2x - 5y = 1$. Since it is the second last item in the test, most examinees had no chance to reach this item (percentage of omit is 37.6%). Therefore, it has lowered the observed proportion correct.

<Figure 5.1>

5.13 The National Curriculum consists of 4 attainment targets, namely "Using & Applying Mathematics", "Number & Algebra", "Shape, Space & Measures" and

"Data Handling". However, the number of items in the ITDA Mathematical Test is far from adequate to cover all relevant knowledge and skills in the targets and the items only focus on the "Number & Algebra" and "Shape, Space & Measures". The test, therefore, has low content validity. It is unavoidable as the ITDA was not developed for U.K. students. In fact, it is a general selection test for universities around the world so it has to be fairly "curriculum free". Nevertheless, whether people have been taught the curriculum will have a considerable impact on their performance. Therefore, this inadequacy might affect the performance of examinees and hence lead to violation of some assumptions. For example, examinees who were good in "Data Handling" but weak in "Shape, Space & Measures" would get poorer results and their abilities would be under-estimated. Moreover, the examinees needed more time to think and hence the test would appear to be "more speeded". They would also be forced to guess hence introducing a relatively larger guessing factor. On the other hand, the abilities of examinees who were good in "Shape, Space & Measures" but weak in "Data handling" would be over-estimated.

Inappropriate Model

5.14 It might be true that the Rasch model and the data are mis-fitted. The Rasch model is a theoretical ideal and incorporates only the general features of the data (item difficulty and non-linearity of raw scores). To incorporate the sample sensitive features of the data (discrimination and guessing), 3 parameter logistic (3PL) model may be used because it has taken the item discrimination index and even guessing factor into consideration. However, the inclusion of the additional parameters (item discrimination and guessing level) requires larger samples of examinees and longer tests to obtain satisfactory item and ability

estimates (Slinde & Linn, 1979, p.451; Goldman & Raju, 1986, p.17). For example, Lord (1980) suggested that sample sizes in excess of 1000 subjects and tests with more than 50 items are required to adequately estimate the item discrimination parameter in the 3PL model. Sample size was not a problem in this study, but the number of items was. There were only 35 items in the ITDA Mathematical Test and 22 and 13 items respectively in the "Number & Algebra" and "Shape, Space & Measures" content categories (also 15 and 20 items respectively in the regular and comparative item form categories) in Bejar's analysis method for checking the assumption of unidimensionality. The corresponding numbers of items in Test-24 were even less. Therefore, the numbers of items were far below the number suggested by Lord.

5.15 Undoubtedly, the multi-parameter models have a greater flexibility but at the same time there are technical problems of using it (Goldstein, 1979, p.215). The program, LOGIST, usually used for 2PL and 3PL models, has been shown to generate estimated parameters with large biases (cited in Fisher, 1994, p.51). Ree & Jensen viewed that guessing cannot be consistently estimated (cited in Wainer & Wright, 1980, p.374). Lord also note that the 3PL model usually does not converge properly and the estimated value of discrimination is likely to increase without limit (cited in Fisher, 1994, pp.51-52). Wright (1977, pp.103-104) showed that "estimates of item discrimination ... drift off to infinity one by one" and he suggested treating the variation in item discrimination through "supervision" rather than estimation. Moreover, it is not uncommon in most 3PL analyses to have observations containing negative information[22], that

[22] By the definition of information (footnote 7 on page 27), it is very clear that the information should be

is, with less certainty about the ability of the person relative to the difficulty of the item than before the observation was made, even when the data fit the model (Wright, 1996, p.504). For example, there will be two implications when a person succeeds on an item in an adaptive test if a 3PL model is used. First, the item is easy to the person (so that a harder item should be provided next). Second, the item is hard to the person so that he/she has to guess and it so happens that he/she has made a lucky guess (so that an easier item should be provided next). Of course, this problem arises mainly from multiple-choice items (in which guessing is a more prominent factor) and the quandary could be avoided to an extent by using constructed responses. However, it will never happen in Rasch model as the guessing factor is assumed to be zero. Disruptions in the measurement process are inevitable but as Fisher (1994, p.62) commented, it is far more productive to locate and interpret the departures from assumptions after they occur than to try to include them as elements in a model of an already very complicated situation.

5.16 The approach of adopting more parameters in doing the estimation is an everlasting process. The discrimination index and guessing factor are only two other popular parameters to be included in the analysis of responses to explain the item sensitivity. Along the same line, should we include parameters to explain the person sensitivity? In fact, disturbances like (i) carelessness (leading to the overestimation of the item difficulty as items are wrongly responded to carelessly); (ii) practice (examinees may require several items to warm up and items affected seem more difficult); (iii) examinees' personality

positive and the greater its value, the smaller will be the standard error of measurement. Therefore, in real practice, items with large standard errors of measurement may be discarded.

(some examinees intend to guess but some never); (iv) examination anxiety (some examinees are very anxious about their performance in tests, but some are not); (v) level of motivation (examinees getting inadequate motivation to respond to the items may just randomly guess), etc. also affect the performance of examinees in a test. Although the Rasch model is robust, to an extent, to such aberrations, they make the estimation procedures both biased and inefficient (Wainer & Wright, 1980, p.374). On the other hand, these disturbances are rarely included in IRT models as the results will be very complicated. Moreover, all the disturbances represent some form of multidimensionality and would violate any model that assumes unidimensionality (Mead, 1976, p.11). Since these disturbances often change the slope of the ICC, any model that includes item discrimination as a parameter would appear to fit the data. Therefore, we have come to an absurd case that we get the data which pass the test of fit but may have in fact violated the model's assumptions. Fitting such a general model may lead to the loss of the desirable measurement properties of the Rasch model and mislead ourselves about the true nature of the variable (p.11).

5.17 Since the ITDA Mathematical Test is used to provide a baseline for students joining the ALIS project, the Rasch model in this study can be considered as an instrument used for solving a practical measurement problem. Thus, the main objective is whether the model can assist to achieve the solution. According to Gustafsson (1980, p.226), "deviations from the model do not necessarily jeopardize applications, ..., the estimates of ability are quite robust against deviations from the model". Therefore, despite the poor fit (particularly for extreme abilities of examinees), there is no point to give up this model.

5.18 A rather simple strategy to obtain fit of ITDA data to the Rasch model is excluding unfitted items from the test. The difficulty parameters of the remaining items are re-calibrated with the item fit re-tested. The process is repeated, excluding more items until a reasonable overall fit is achieved. However, if items are dropped from a test, that test would no longer match the test specification, particularly if the proportion of discarded items is large, and has lost content validity (Divgi, 1986, p.295; Phillips, 1986, p.107). Gustafsson (1980, p.230) holds similar view that misfitting items should not be routinely excluded from a test to obtain fit to the model as goodness-of-fit cannot replace subject matter knowledge. Therefore, it is desirable to identify the likely causes of the poor fit (like speededness, guessing, etc.) and take proper actions to remove those threats against the model. For example, speededness in this study may be one of the possible causes of the poor fit.

Comparability ot the CTT Method and the Rasch Method

5.19 In this study, the comparability of the CTT method and the Rasch model gives a very interesting result. The findings failed to discredit the CTT framework with regard to its alleged inability to produce person-free item difficulty estimates and item-free person ability estimates. The findings simply revealed that the two measurement frameworks produced very similar item and person statistics both in terms of comparability and invariance. Nevertheless, the present study has its limitations that may lower the validity of the findings. First, 75% of examinees got a raw total score of 17 or below (as reflected in para.3.2 where the High-ability Group and Low-ability Group were defined). The test score distribution has a positive skewness of 1.692 and shows a strong

floor effect which suggests that many items tended to be hard to most examinees. The difficulty of some items might be due to the speededness of the test as a large percentage of examinees could not complete the test and all omits (including not-reached responses) were scored as zero. These items, particularly those at the end of the test, would be overestimated in terms of difficulty. Therefore, it would be desirable in future studies to use data from a non-speeded test which involves items varying in item difficulty (to avoid both floor and ceiling effects).

5.20 The second shortcoming of the comparison is the limited item pool used. The examinee pool is quite adequate in the sense that a variety of different samples can be drawn for investigation, but the same cannot be said about the item pool. There was only one test in the study, namely the ITDA Mathematical Test, although the test was re-grouped to different item groups. Therefore, the comparison of the two frameworks could be replicated with a larger test item pool which is more diverse in terms of item characteristics. Items can then be sampled from the pool to study the behaviours of CTT and Rasch item statistics under different conditions of item characteristics.

5.21 The last 11 items of the ITDA Mathematical Test were deleted to take the factor of speededness into account and the data of the resulting test (i.e. Test-24) were re-analyzed. The findings, however, revealed no significant improvement. The possible explanations are:

(a) examinees were assumed to respond to items in serial order, but in fact some of them did not;

(b) the time pressure and the resulting psychological effect due to

speededness like carelessness, anxiety, speed-accuracy trade-off attitude, etc. had been ignored.

Recommendations

5.22 In the present study, the feature of item parameter invariance was studied through random samples from the Whole Group, gender groups and ability groups. Since the violation of non-speededness has a more serious influence to the examinees of low abilities (for example, encouraging them to guess), the feature of invariance might have been confounded. Therefore, in future studies, other subgroups of special interest in the examinee population could be identified and analyzed for the feature of invariance. For example, it is meaningful to consider examinees with different socio-economic status, from different ethnic groups or geographic regions. The relevant particulars should, of course, be collected from the examinees sitting the test for identifying the subgroups of interest.

5.23 With the advance of information technology, it is possible to deliver the ITDA test electronically or over the internet in some way. The response time of each item can be recorded (i.e. the examinee's speed in responding to each individual item can be tracked). In this way, it is possible to get a measure of speed without putting a constraint on the total time available to the examinees for the full test and a measurement model to take speed into account can be adopted. For example, the Rasch Response Time Model proposed by Roskam (cited in Roskam, 1997, p.193) has integrated the response time and correctness as follows:

$$P_{xi}(\theta_x, b_i, \tau_{xi}) = \frac{1}{1 + e^{-(\theta_x + \tau_{xi} - b_i)}}$$

where τ_{xi} is the natural logarithm of the response time of examinee x on item i. When the response times are observed, the estimation procedure of the item and person parameters are basically the same as in the dichotomous Rasch model.

5.24 Marginal Maximum Likelihood and Bayesian estimation procedures could be adopted instead of the Unconditional Maximum Likelihood procedure to take into account the estimates of zero and perfect scores. However, there is no viable method to improve the estimation with regard to omits and not-reached items. What could be done is to reduce the effect of speededness and hence reduce omits and not-reached items. Slight modifications to the ITDA Mathematical Test are therefore recommended: Due to the possible influence of the quantitative comparison items on the low-ability examinees, items 16 – 35 of the ITDA test could be replaced by items of regular form to avoid the possible effect. If the item form could not be changed, more time should be allowed for the examinees for reading the instructions carefully. If possible, one or two simple items may be added for practice. In particular, items 16 – 30 are crowded in the same page. It could be spread uniformly across two pages to make it more spacious and easier to read. The items may also be presented in an increasing order of difficulty (see Table 4:8) so that examinees get motivation to continue the test and hence reduce omits. If an examinee reaches items which are too difficult to respond to, he or she may stop as it could be concluded that trying any of the remaining items is a waste of time. In this way, the bias imputable to scoring the not-reached responses as zero might be reduced.

5.25 From the discussion above, the Rasch model could be used to calibrate the test items as it is a theoretical ideal. On the other hand, improved estimation procedures could be applied for estimating examinee abilities. Those proposed by Wainer & Wright (1980), namely jackknife and AMT-robustified jackknife could be considered. Through a Monte Carlo simulation study, Wainer & Wright found that the jackknife and the AMT-robustified jackknife estimators yielded better estimates of ability than the maximum likelihood estimator for tests with 40 or fewer items and the AMT-jackknife was even better when dealing with guessing. Although the simulation had assumed the availability of item difficulties and only estimated abilities, "some of the techniques ... will be of some use in the estimation of item difficulties" (Wainer & Wright, 1980, p.374). Another modified version of the Rasch model introduced by Waller (cited in Slinde & Linn, 1979, pp.451-452) may also be useful. This modified procedure is an application of the Abilities Removing Random Guessing Model which assumes that examinees randomly guess on those items that are too difficult for them. Thus, this procedure involves removing from the estimation procedure the response of an examinee to an item estimated to be very difficult for him/her. It could be programmed into the marking procedures.

APPENDICES

Appendix 1

Proving that the Number Correct Score is a Sufficient Statistic for θ

The proof below is abridged from Lord (1980, p.63):

Equation (2.10) in Chapter 2 can be rearranged as

$$L_x(u_{x1}, u_{x2}, \ldots, u_{xn}|\theta_x, (b_i)) = \prod_{i=1}^{n}\left(\frac{P_{xi}}{Q_{xi}}\right)^{u_{xi}} Q_{xi}$$

$$= \prod_{i=1}^{n}\left(\frac{P_{xi}}{Q_{xi}}\right)^{u_{xi}} \prod_{i=1}^{n} Q_{xi}$$

$$= \prod_{i=1}^{n} e^{D(\theta_x - b_i)u_{xi}} \prod_{i=1}^{n} Q_{xi}$$

$$= e^{D\theta_x r_x} e^{-D\sum_{i=1}^{n} b_i u_{xi}} \prod_{i=1}^{n} Q_{xi} \qquad \text{(A1.1)}$$

where $r_x = \sum_{i=1}^{n} u_{xi}$. Since P(A and B) = P(A) P(B|A),

$$P(r_x \text{ and } (u_{xi})|\theta_x) = P(r_x|\theta_x)P((u_{xi})|r_x, \theta_x)$$

or
$$P((u_{xi})|r_x, \theta_x) = \frac{P(r_x \text{ and } (u_{xi})|\theta_x)}{P(r_x|\theta_x)} \qquad \text{(A1.2)}$$

where (u_{xi}) is the 1 × n vector with elements u_{xi} (i = 1, 2, 3, … , n) for the examinee x.

As r_x depends on (u_{xi}), (A1.2) can be simplified to

$$P((u_{xi})|r_x, \theta_x) = \frac{P((u_{xi})|\theta_x)}{\sum_{(u_{xi})|r_x} P((u_{xi})|\theta_x)} \qquad \text{(A1.3)}$$

where the summation is over all (u_{xi}) for which $\sum_{i=1}^{n} u_{xi} = r_x$.

Substituting (A1.1) into (A1.3),

$$P((u_{xi})|r_x,\theta_x) = \frac{e^{D\theta_x r_x} e^{-D\sum_{i=1}^{n} b_i u_{xi}} \prod_{i=1}^{n} Q_{xi}}{\sum_{(u_{xi})|r_x} e^{D\theta_x r_x} e^{-D\sum_{i=1}^{n} b_i u_{xi}} \prod_{i=1}^{n} Q_{xi}}$$

$$= \frac{e^{-D\sum_{i=1}^{n} b_i u_{xi}}}{\sum_{(u_{xi})|r_x} e^{-D\sum_{i=1}^{n} b_i u_{xi}}}. \qquad \text{(A1.4)}$$

Clearly, the right hand side of (A1.4) is independent of θ. Therefore, the number correct score, $r_x = \sum_{i=1}^{n} u_{xi}$, is a sufficient statistic for the ability θ.

Appendix 2

Finding the Slope β and Intercept α of the Principal Axis

Let the equation of the principal axis of the plot of the two sets of difficulty estimates be $B_{1i} = \beta B_{2i} + \alpha$ where B_{1i} and B_{2i} are the content-area-based and total-test-based estimates respectively, α and β are the corresponding "B_1-intercept" and slope of the line. For simplicity, the subscript i is dropped so that

$$B_1 = \beta B_2 + \alpha . \qquad \text{(A2.1)}$$

To obtain the distance from any point (B_2, B_1) to the principal axis, the two axes are first rotated about the origin by an angle θ (where tan θ = β) and then linearly transformed so that the point (α, 0) becomes the new origin. The transformation is accomplished by the following equation:

$$\begin{pmatrix} B_2' \\ B_1' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} B_2 \\ B_1 \end{pmatrix} - \begin{pmatrix} \alpha\sin\theta \\ \alpha\cos\theta \end{pmatrix} \qquad \text{(A2.2)}$$

where B_1' and B_2' are the respective new values of B_1 and B_2 under the new system of axes. Clearly, B_1' is the distance of any point to the principal axis in the plot where

$$B_1' = -B_2 \sin\theta + B_1 \cos\theta - \alpha\cos\theta . \qquad \text{(A2.3)}$$

If D is the sum of the squared distances of all points in the plot to the principal axis. Then,

$$D = \sum(-B_2 \sin\theta + B_1 \cos\theta - \alpha\cos\theta)^2 \qquad \text{(A2.4)}$$

where the summation is over all estimates, i.e. i = 1 to n. The values of α and β (or θ) which will give a minimum of D can be found by solving the following two equations:

$$\frac{\partial D}{\partial \alpha} = 0 \qquad \text{(A2.5)}$$

and $$\frac{\partial D}{\partial \theta} = 0 \qquad \text{(A2.6)}$$

From (A2.5), $$\frac{\partial D}{\partial \alpha} = -2\sum(-B_2 \sin\theta + B_1 \cos\theta - \alpha\cos\theta)\cos\theta = 0$$

$$\sum \alpha \cos\theta = \sum(B_1 \cos\theta - B_2 \sin\theta)$$

$$n\alpha = \sum B_1 - \tan\theta \sum B_2$$

$$\alpha = \overline{B}_1 - \tan\theta \overline{B}_2 \quad (\text{or } \alpha = \overline{B}_1 - \beta\overline{B}_2) \qquad \text{(A2.7)}$$

From (A2.7), $$\alpha\cos\theta = \overline{B}_1 \cos\theta - \overline{B}_2 \sin\theta \qquad \text{(A2.8)}$$

Substituting (A2.8) into (A2.4) which then becomes,

$$D = \sum\left[(B_1 - \overline{B}_1)\cos\theta - (B_2 - \overline{B}_2)\sin\theta\right]^2$$

$$= \sum[(B_1 - \overline{B}_1)^2 \cos^2\theta + (B_2 - \overline{B}_2)^2 \sin^2\theta$$

$$-2(B_1 - \overline{B}_1)(B_2 - \overline{B}_2)\sin\theta\cos\theta]$$

Therefore, from (A2.6),

$$\frac{\partial D}{\partial \theta} = \sum[-(B_1 - \overline{B}_1)^2 \sin 2\theta + (B_2 - \overline{B}_2)^2 \sin 2\theta$$

$$-2(B_1 - \overline{B}_1)(B_2 - \overline{B}_2)\cos 2\theta] = 0$$

or $$-s_1^{\,2} \sin 2\theta + s_2^{\,2} \sin 2\theta = 2s_{12} \cos 2\theta$$

where $s_1^{\,2}$ and $s_2^{\,2}$ are the variances of the content-area-based and total-test-based estimates respectively, and s_{12} is the covariance between the two sets of estimates. Rearranging,

$$\tan 2\theta = \frac{2s_{12}}{(-s_1^{\,2} + s_2^{\,2})} \quad \text{or} \quad \frac{2\beta}{1-\beta^2} = \frac{2s_{12}}{(-s_1^{\,2} + s_2^{\,2})}$$

$$s_{12}\beta^2 - \beta(s_1^{\,2} - s_2^{\,2}) - s_{12} = 0$$

or $$\beta = \frac{(s_1^{\,2} - s_2^{\,2}) \pm \sqrt{(s_1^{\,2} - s_2^{\,2})^2 + 4s_{12}^{\,2}}}{2s_{12}} \qquad \text{(A2.9)}$$

where the sign in the numerator is chosen so that β is positive.

Appendix 3

List of Figures

Figure 1.1

Three ICCs for the one-parameter logistic model

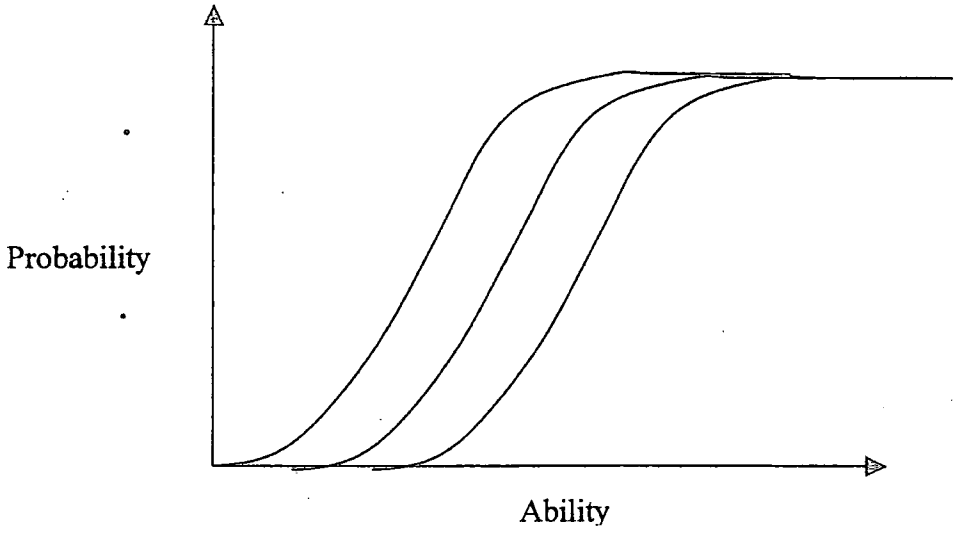


Figure 1.2

Two ICCs for the two-parameter logistic model

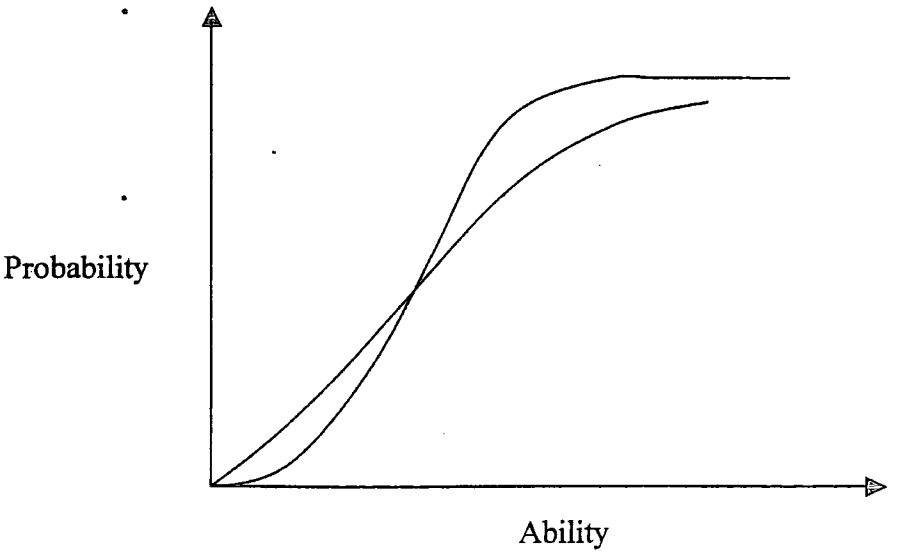


Figure 1.3

Four ICCs for the three-parameter logistic model

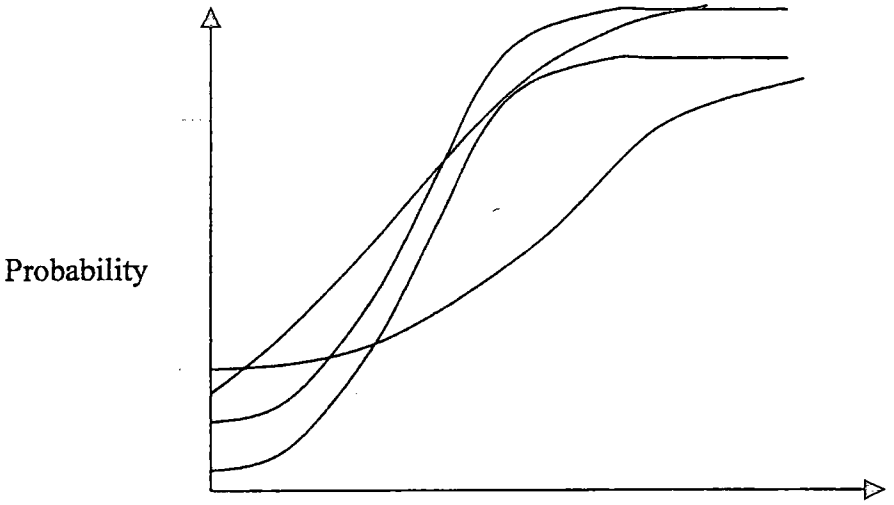


Figure 1.4

Part of the Data File of the Scores of the ITDA Mathematics Items

```
20121852 1101110110010011111011000110010000
20131852 110010x1010100x100101x0000110000x00
20141852 0100100100110001101101100001000xxxxx
10151852 0110100010000001001010110101001000l
20161852 1100000100000000011000101000011010O
10171852 110000101010110100101000001110x10xx
20181852 0100000000101000011101100000100l010
20191852 111010x1000000011011110Ox0110110100
20201852 1100110000110001000011011100000010O
20211852 1111010110000000011100100011000000O
10221852 1101100110000101011011l0x0110000000
20231852 1111000110110001001101111000001l000
20241852 0000100100010111100111100000001011O
10251852 1101x0xx1x10000111111101000101lxxxx
10261852 1111100111000101100101Ox00010010xxx
20271852 1100000111010001011111110101101l010
20281852 1111100100010001101101110011010Oxxx
10291852 111111011101100xxxxxxxxxxxxxxxxxxxx
20301852 101x1101111000010111111010111000000
20311852 1100010000000001101001101011001lll0
10321852 1100100100000010x1110110101x0xxxx0
```

Figure 1.5
Frequency Polygon for the Performance of Examinees
(Skewness = 1.692)

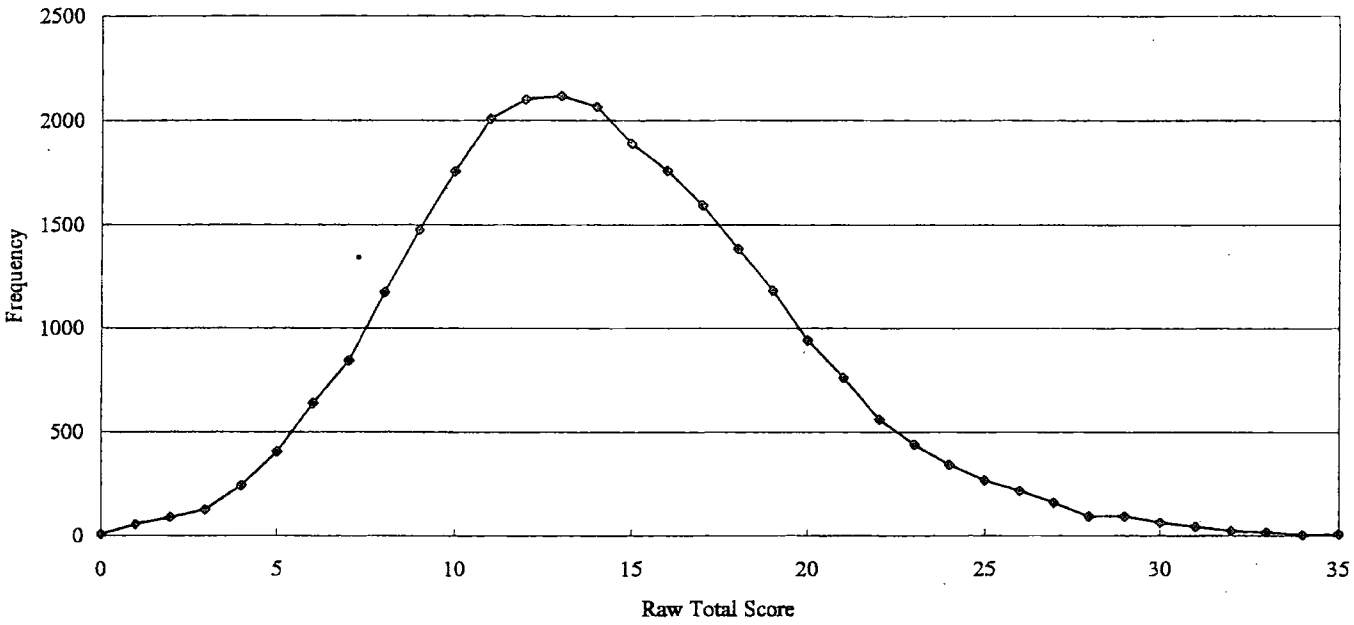


Figure 1.6
Cumulative Frequency Polygon for the Performance of Examinees
(Skewness = 1.692)

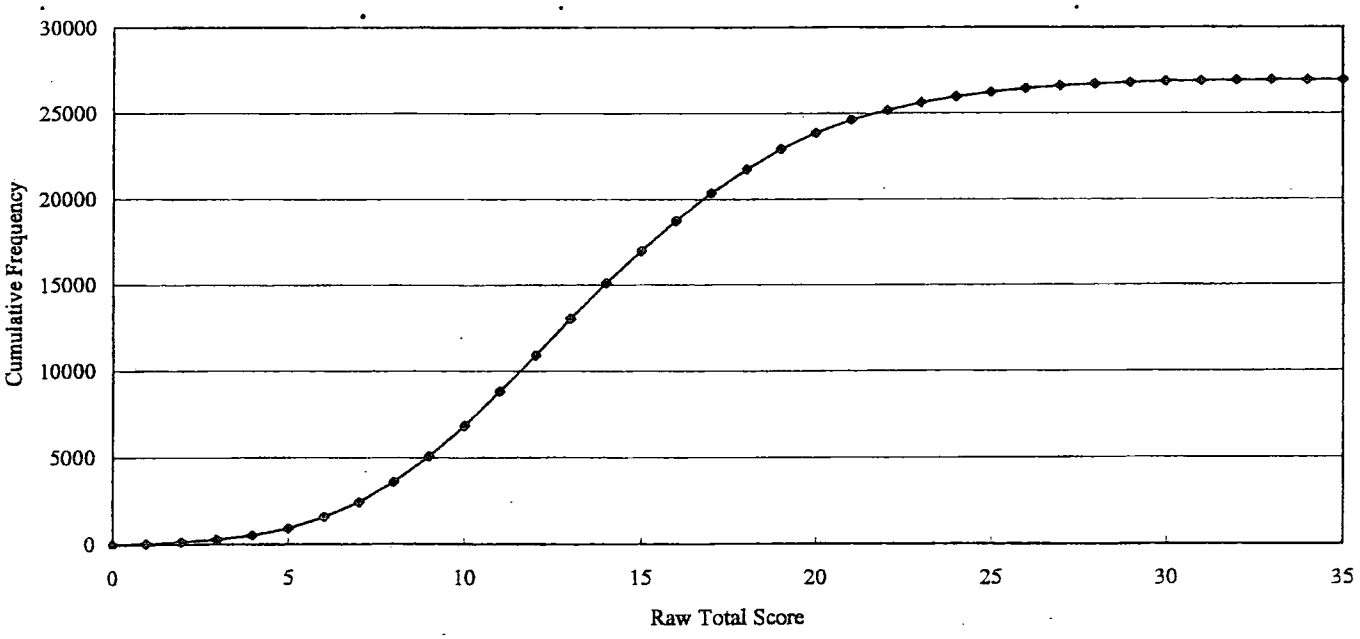
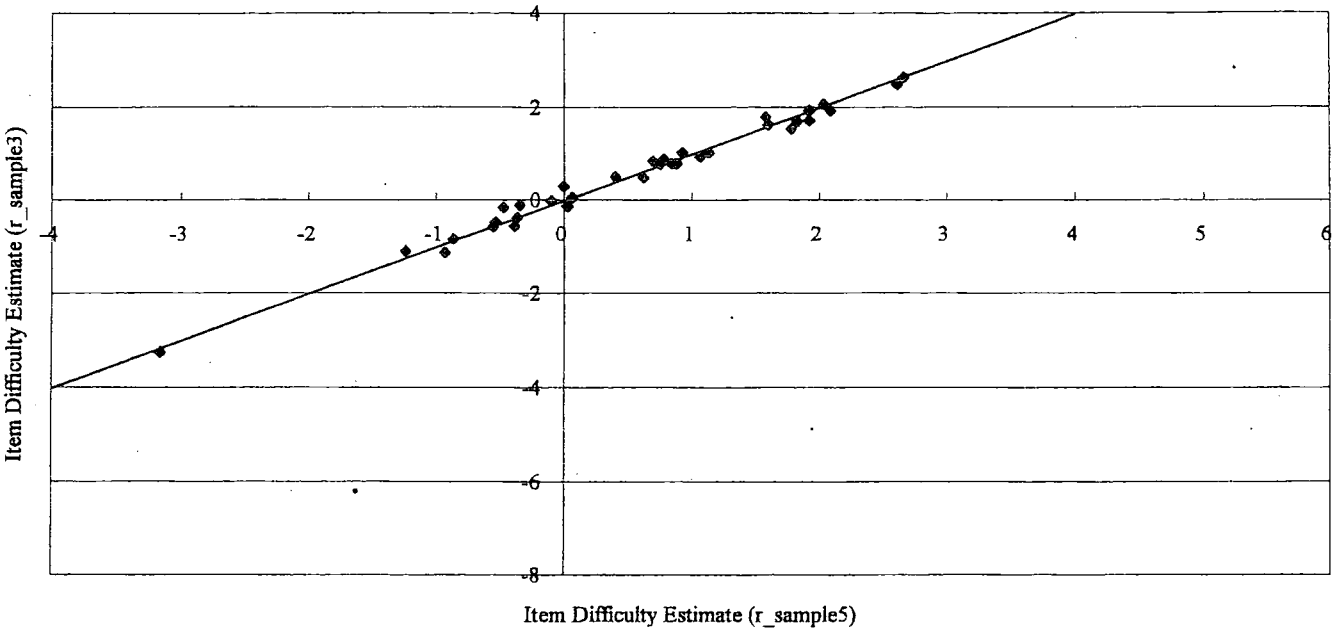


Figure 4.1
Scatterplot of Item Difficulty Estimates (r_sample3 vs. r_sample5)
(a) Test-35 (r = 0.993)



(b) Test-24 (r = 0.990)

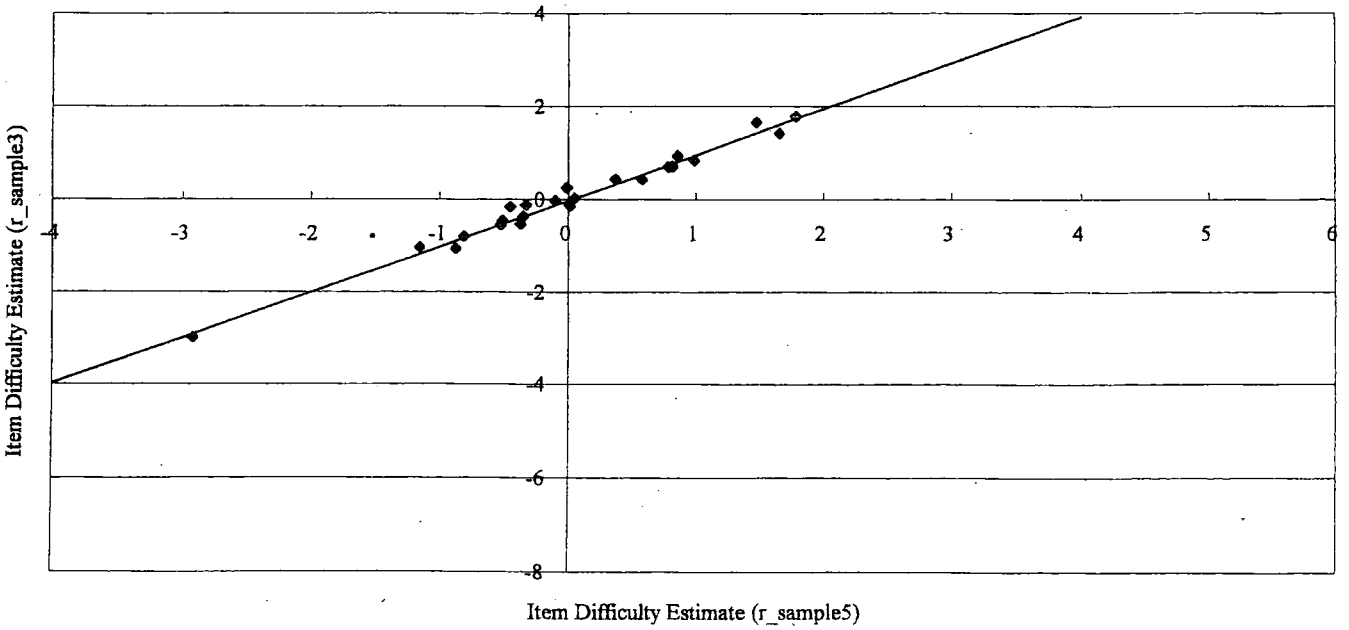
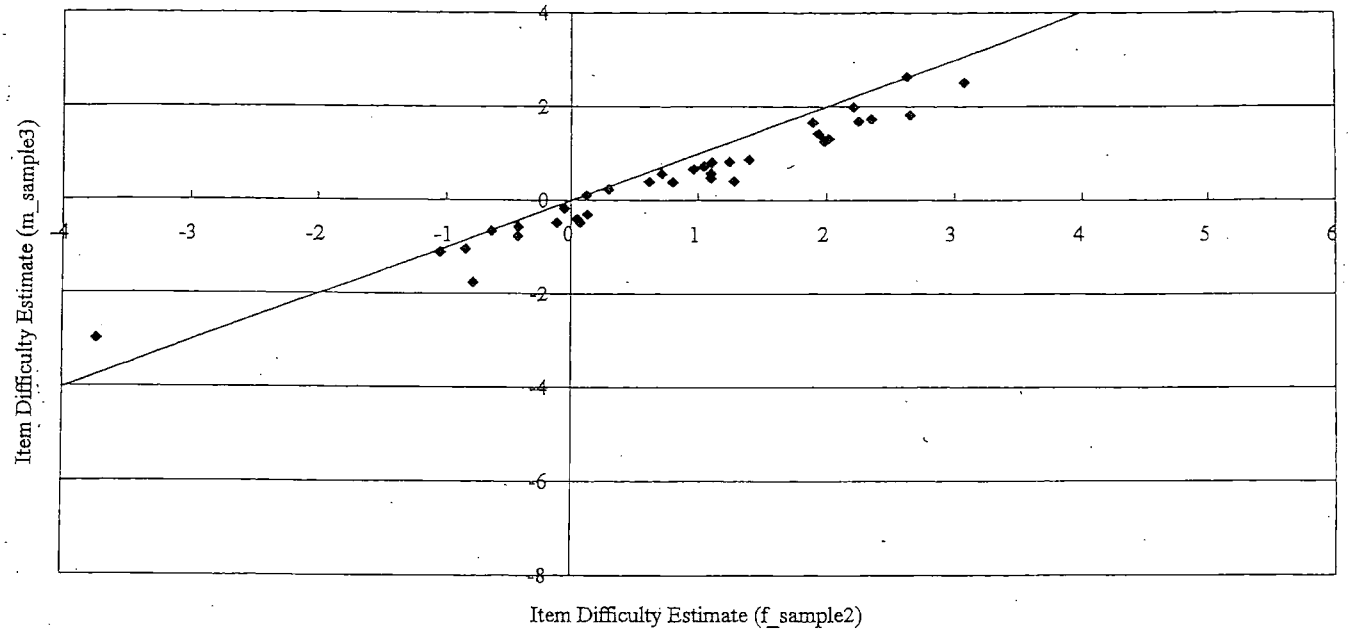


Figure 4.2
Scatterplot of Item Difficulty Estimates (m_sample3 vs. f_sample2)
(a) Test-35 (r = 0.976)



(b) Test-24 (r = 0.974)

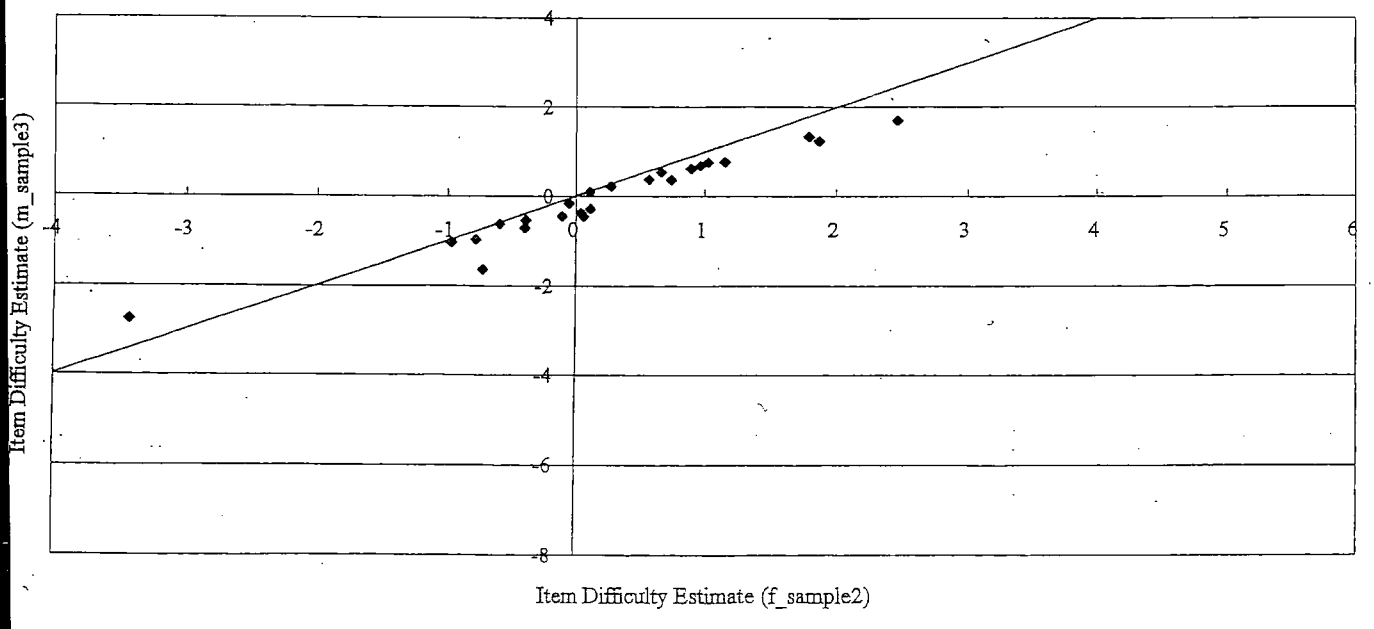
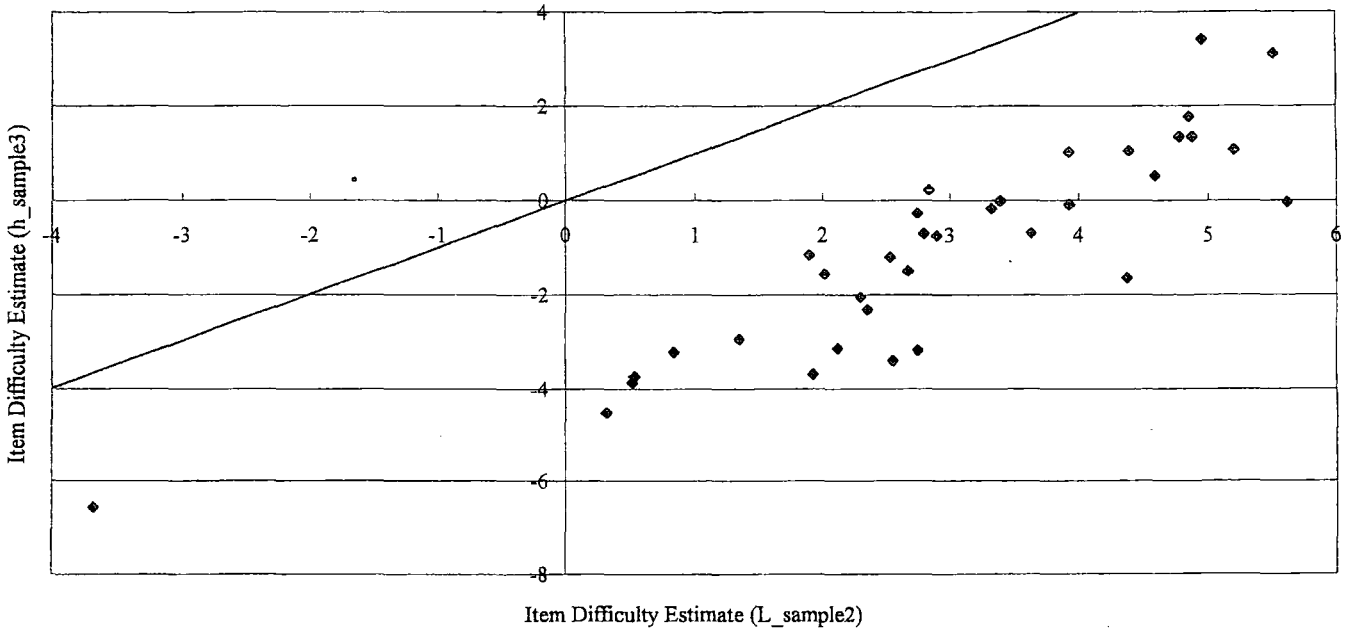


Figure 4.3
Scatterplot of Item Difficulty Estimates (h_sample3 vs. L_sample2)
(a) Test-35 (r = 0.882)



(b) Test-24 (r = 0.872)

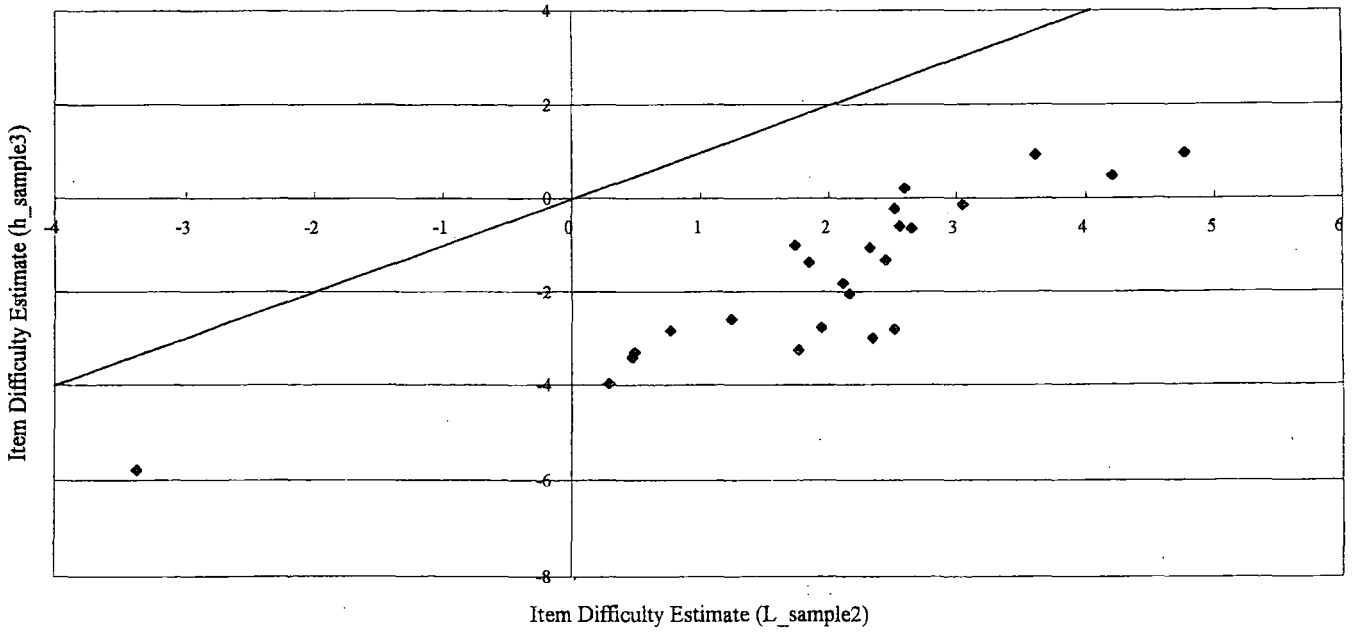


Figure 4.4
Effect Sizes of Item Difficulty Estimates

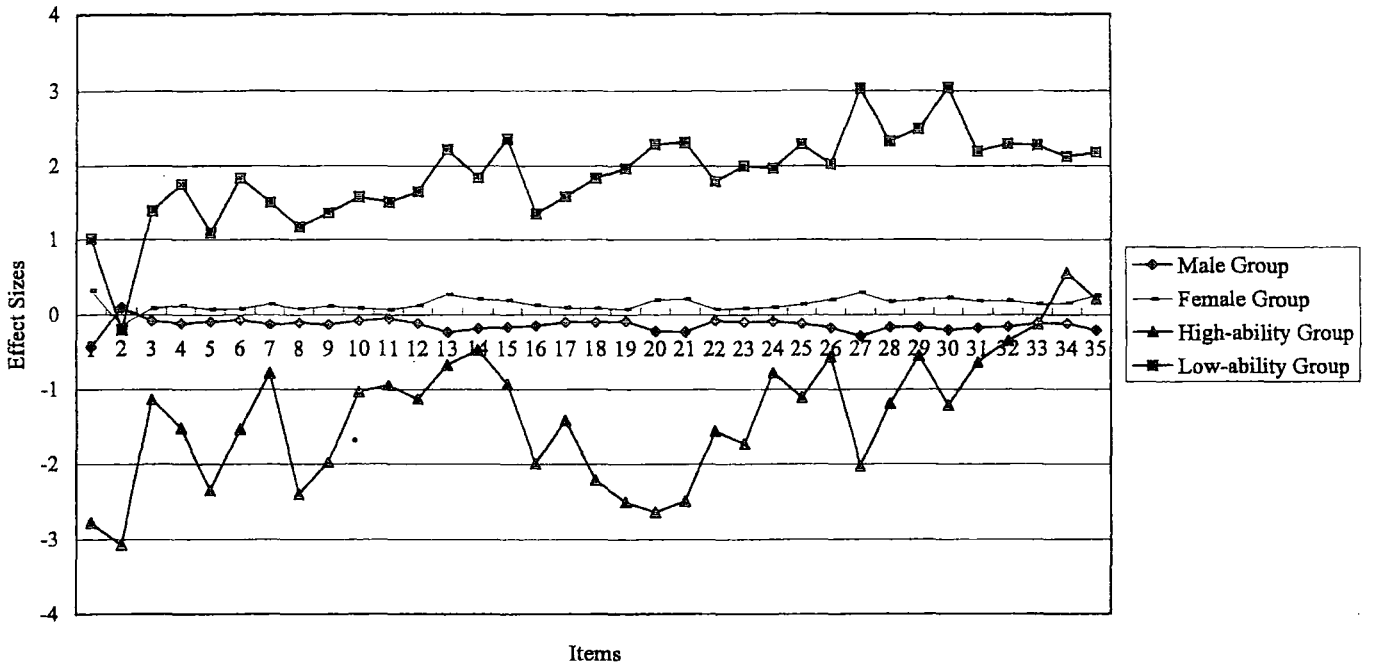
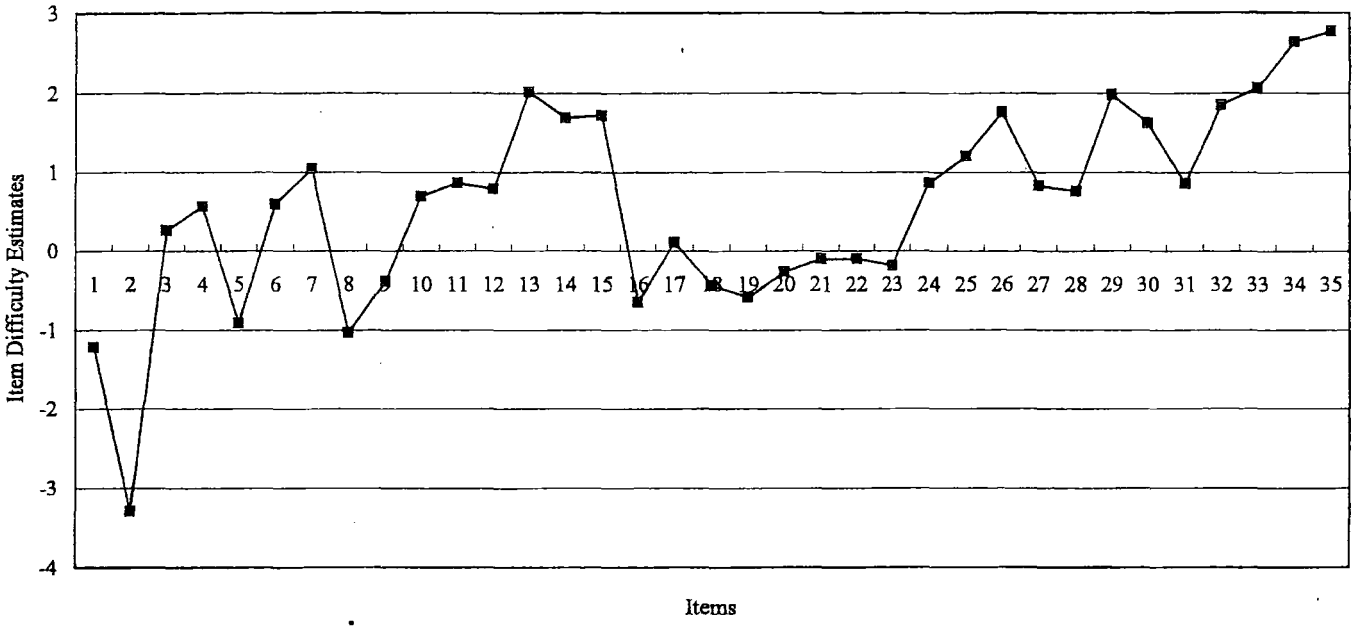


Figure 4.5
Item by Difficulty Estimates
(a) Test-35



(b) Test-24

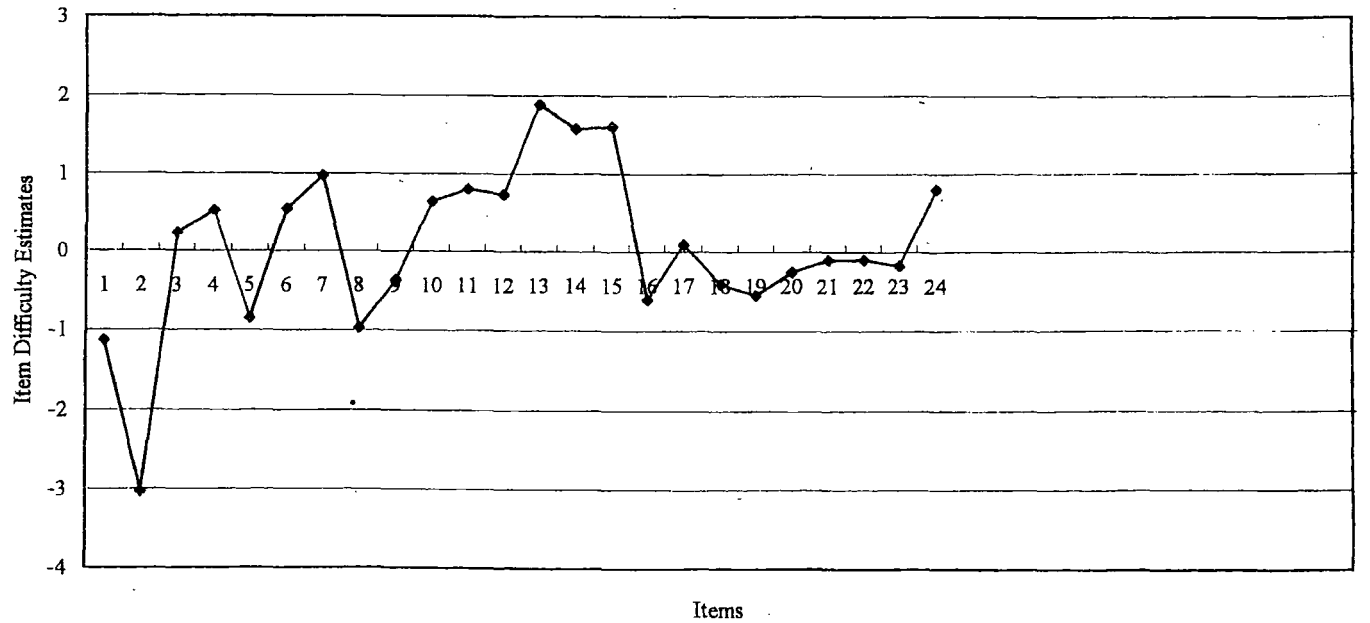
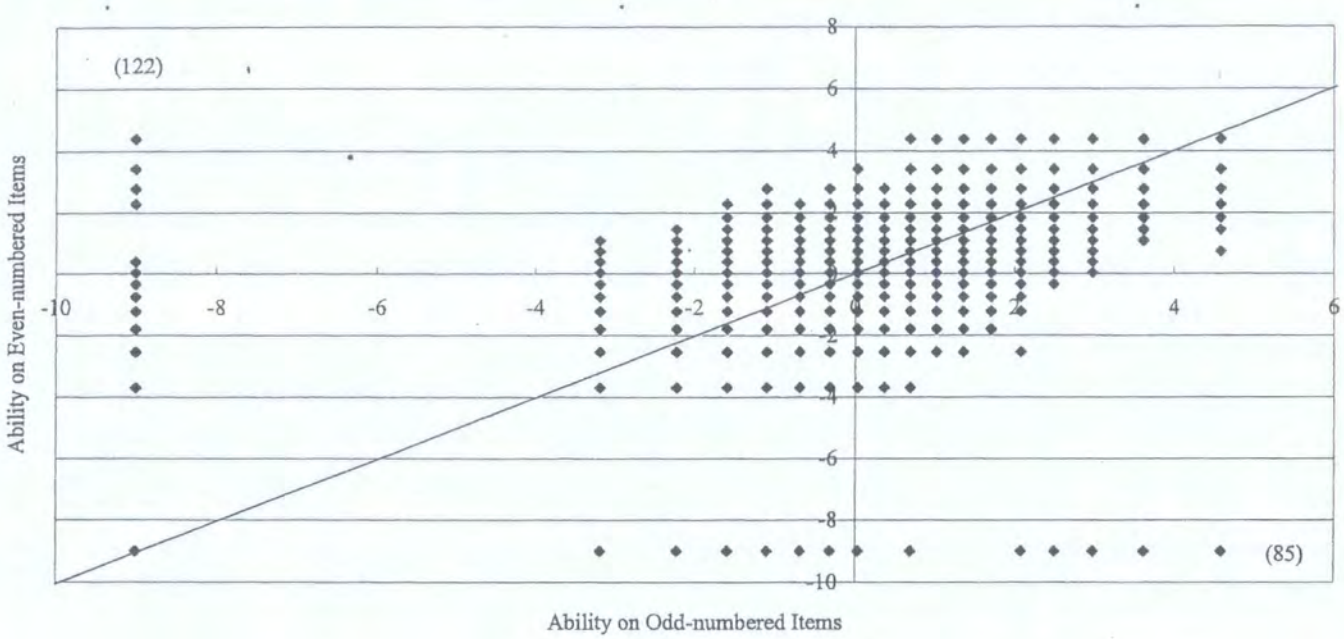


Figure 4.6
Scatterplot of Ability Estimates based on Equivalent Halves
(a) Test-35 (r = 0.580)



(b) Test-24 (r = 0.342)

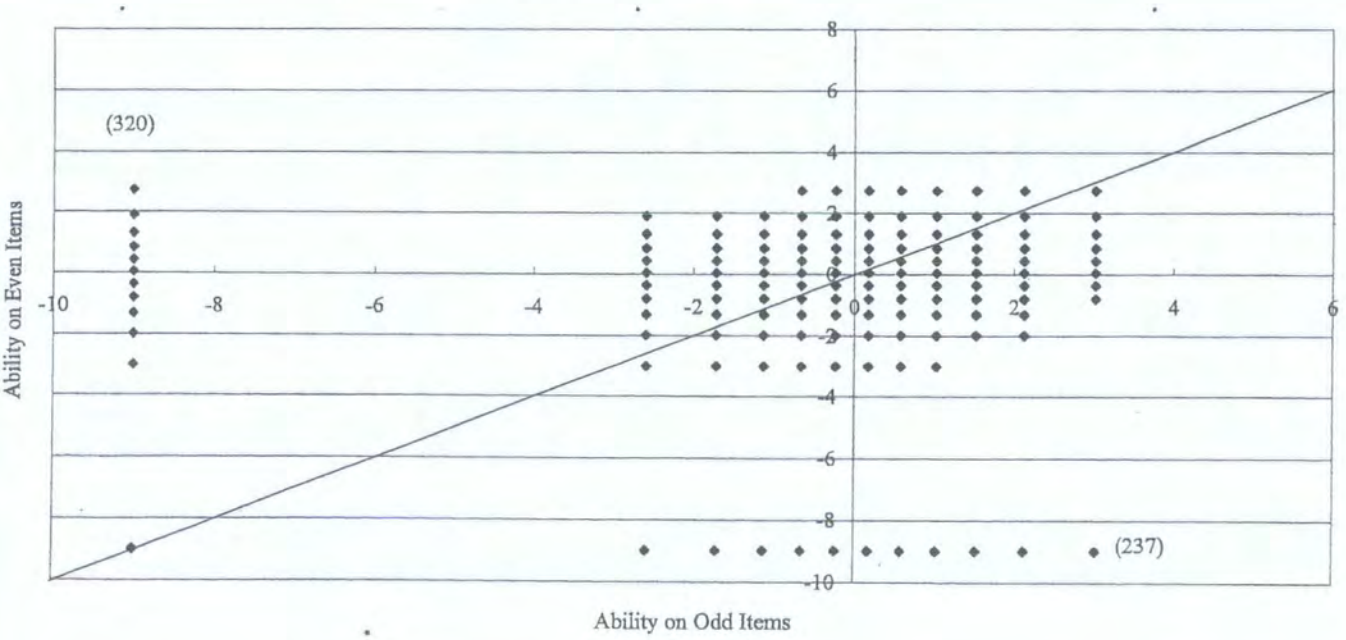
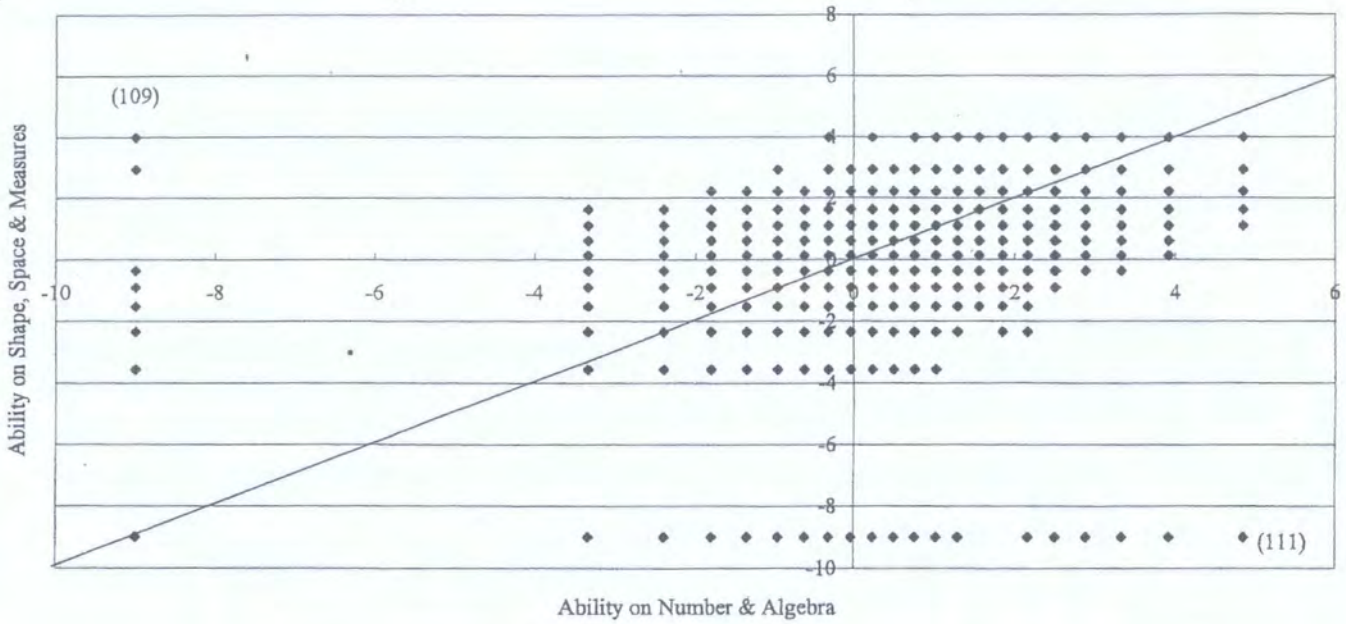


Figure 4.7
Scatterplot of Ability Estimates based on Different Content Categories
(a) Test-35 (r = 0.527)



(b) Test-24 (r = 0.285)

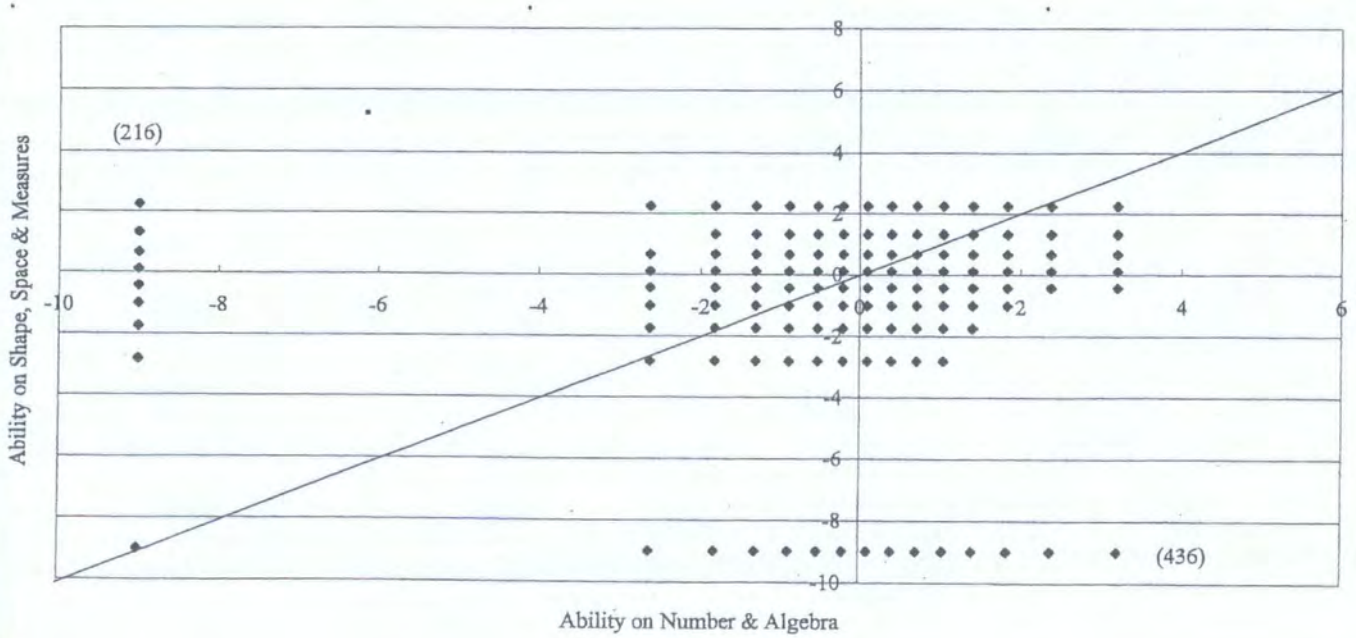
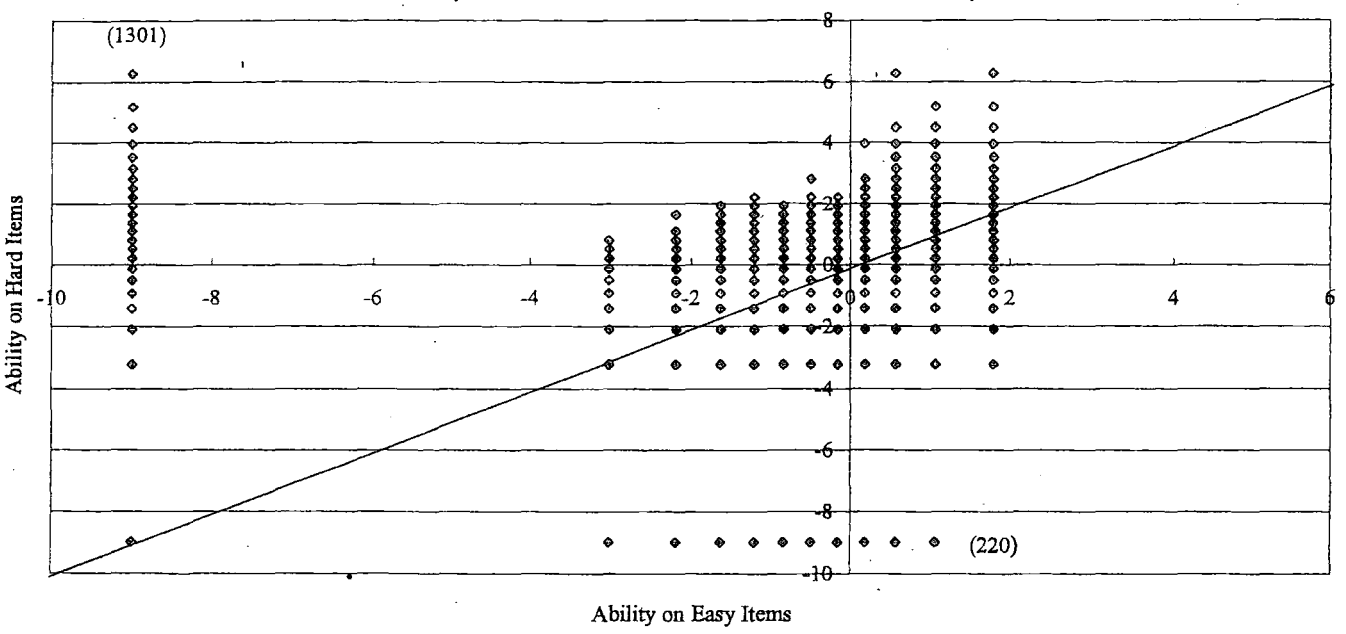


Figure 4.8
Scatterplot of Ability Estimates based on Items of Different Difficulties
(a) Test-35 (r = 0.013)



(b) Test-24 (r = 0.023)

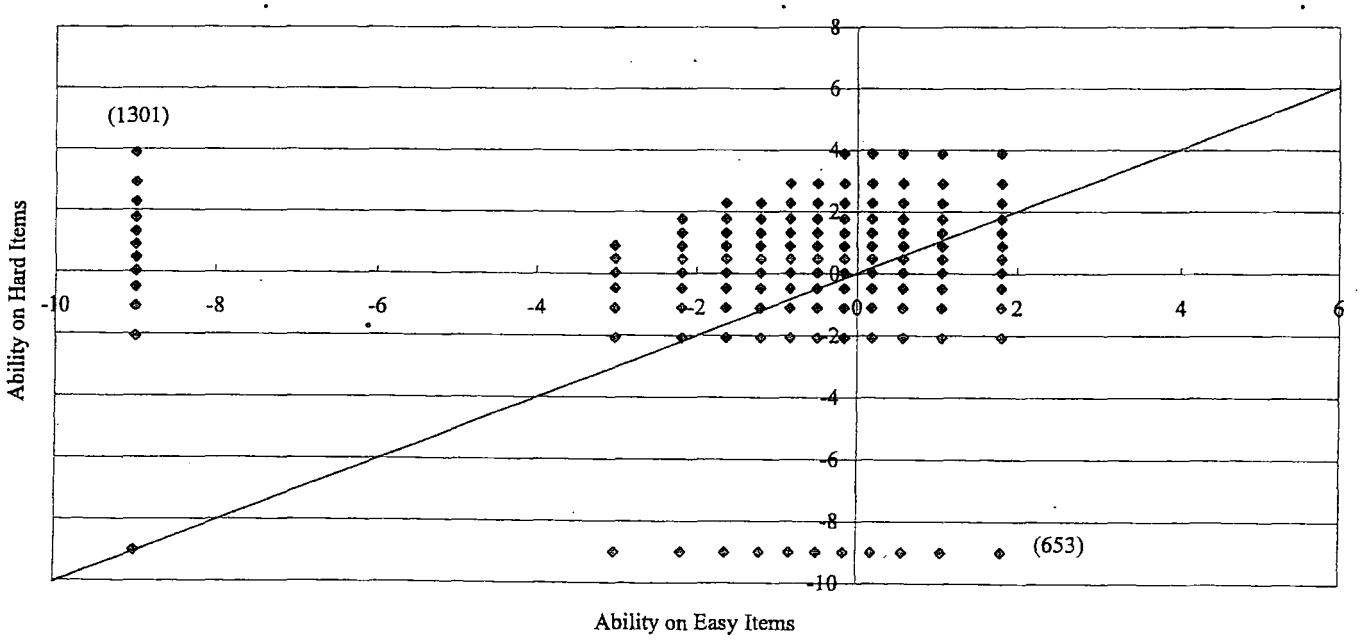
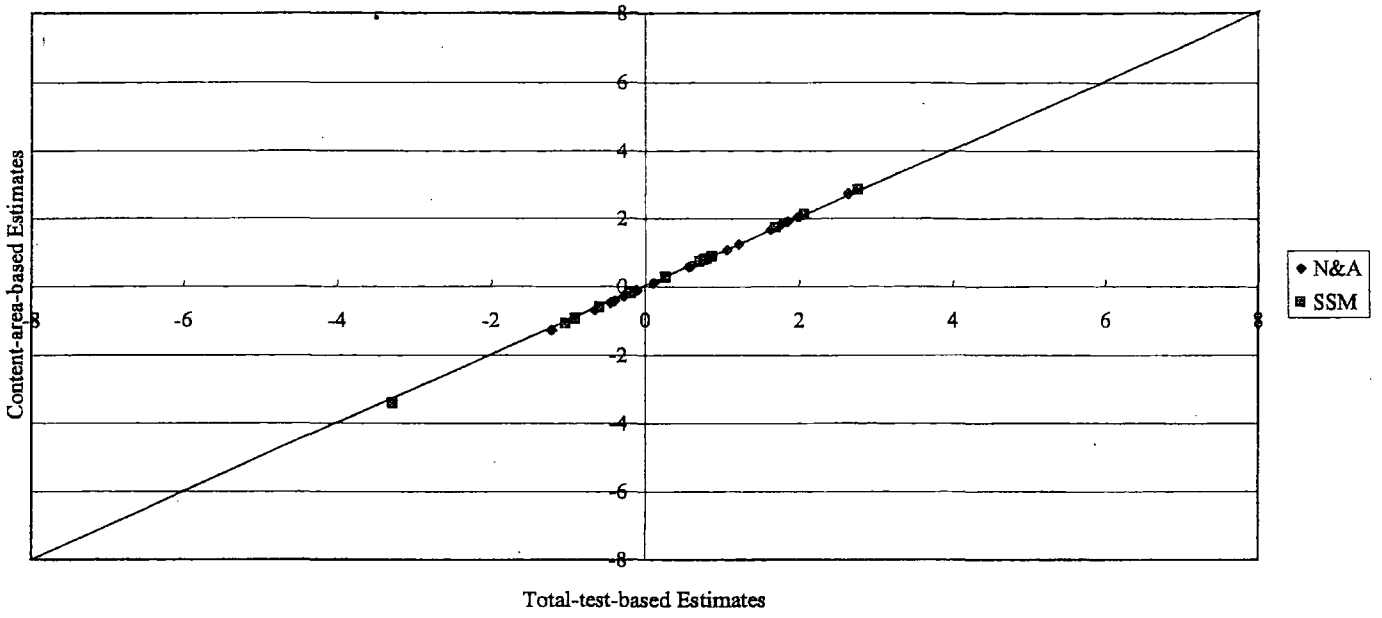


Figure 4.9
Plot of Content-area-based Difficulty Estimates vs.
Total-test-based Estimates for the Whole Group
(a) Test-35 (r = 1.000)



(b) Test-24 (r = 1.000)

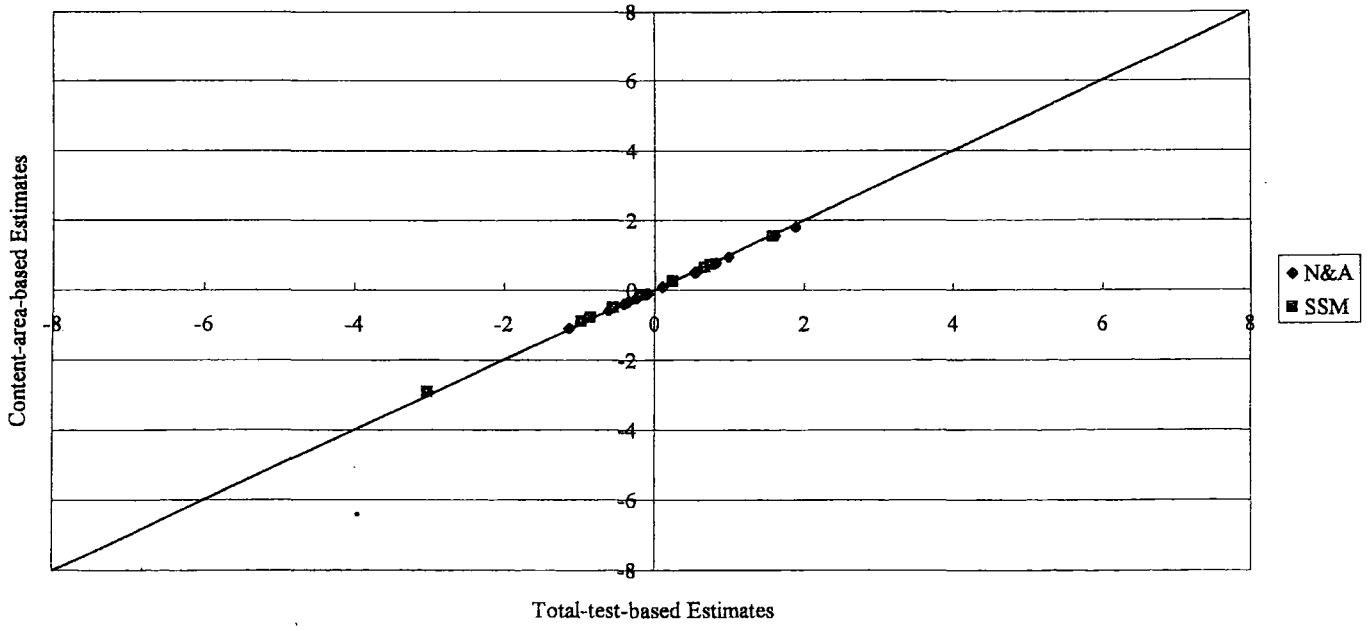
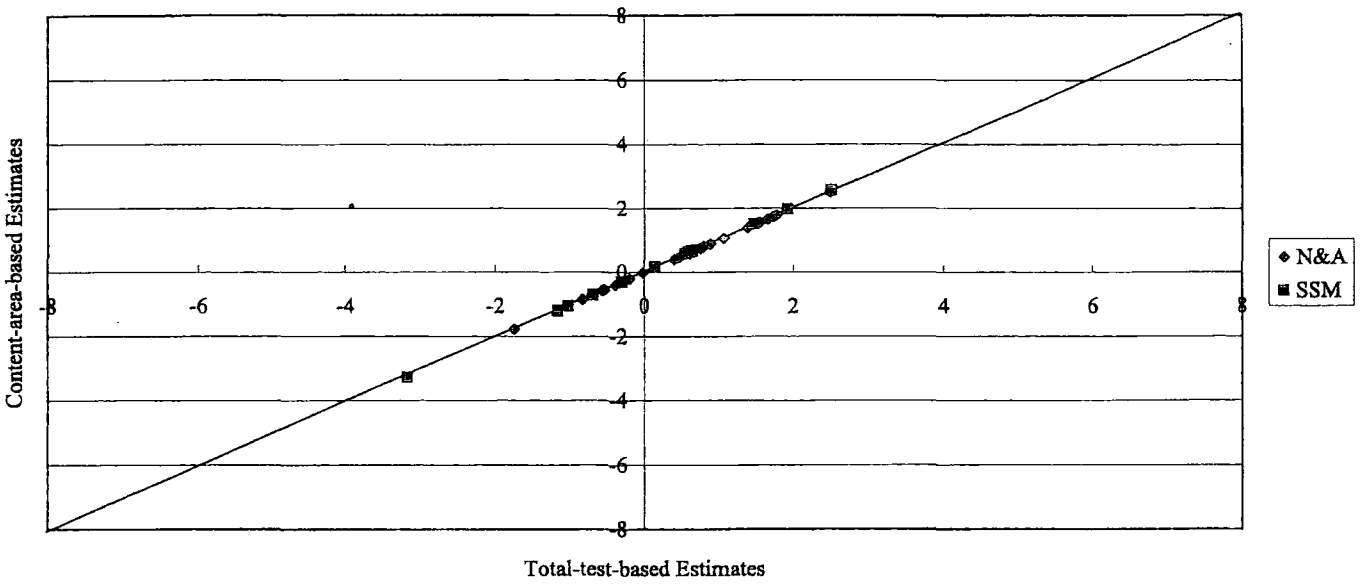


Figure 4.10
Plot of Content-area-based Difficulty Estimates vs.
Total-test-based Estimates for the Male Group
(a) Test-35 (r = 1.000)



(b) Test-24 (r = 1.000)

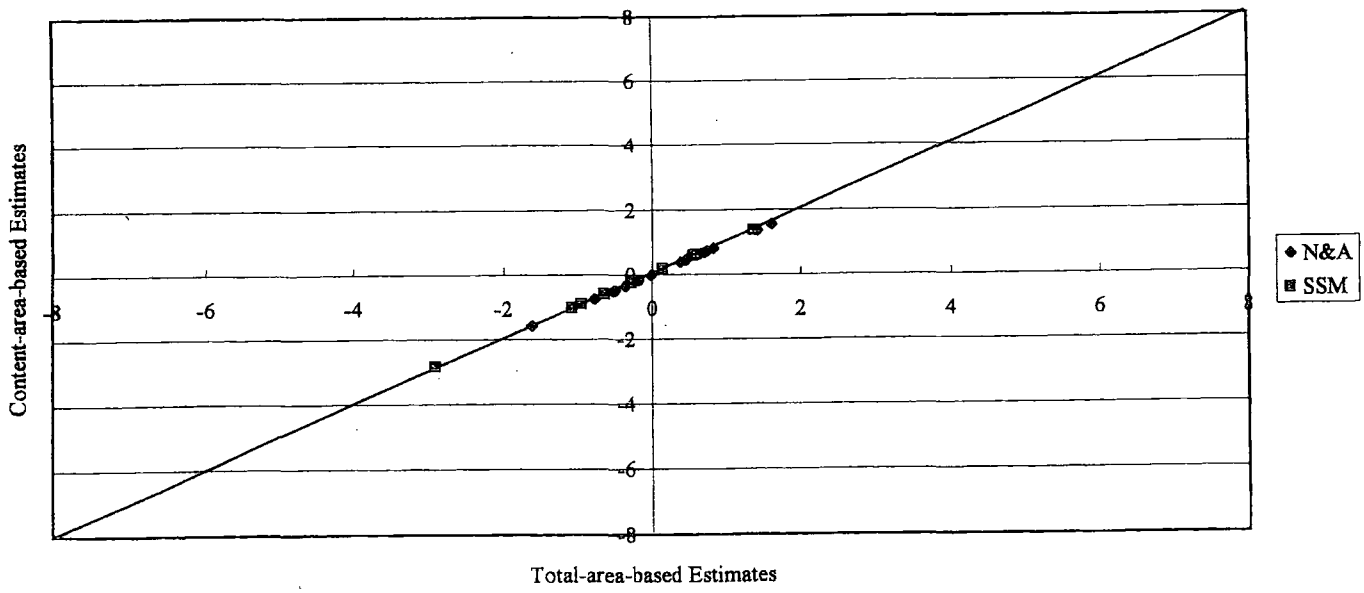
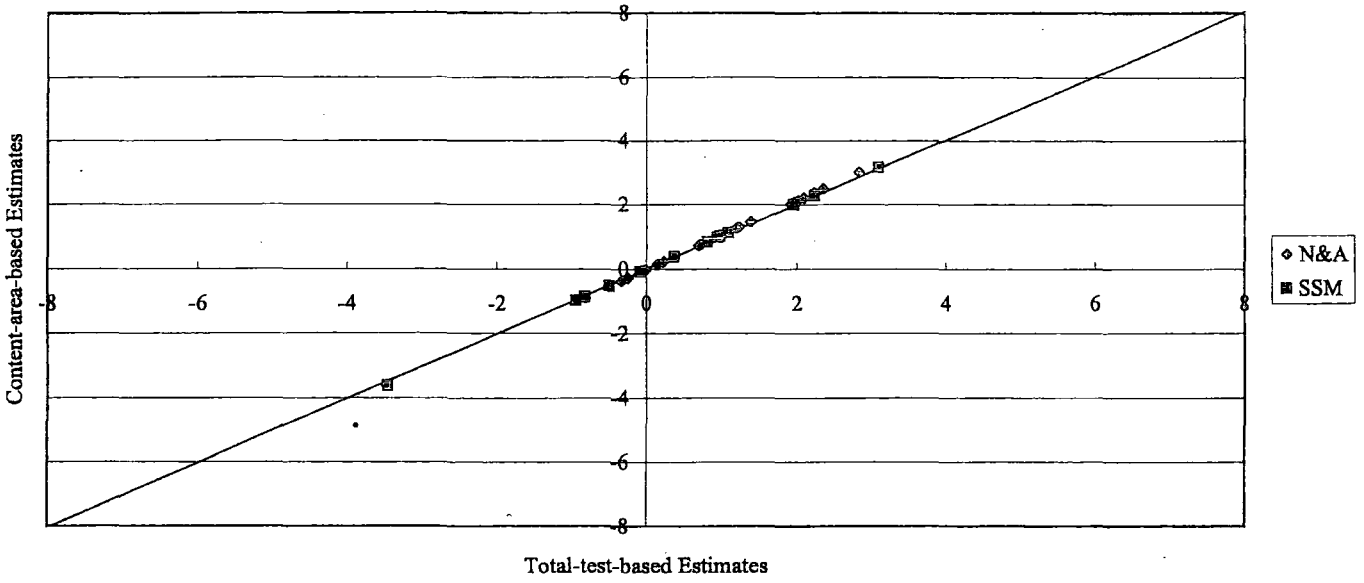


Figure 4.11
Plot of Content-area-based Difficulty Estimates vs.
Total-test-based Estimates for the Female Group
(a) Test-35 (r = 1.000)



(b) Test-24 (r = 1.000)

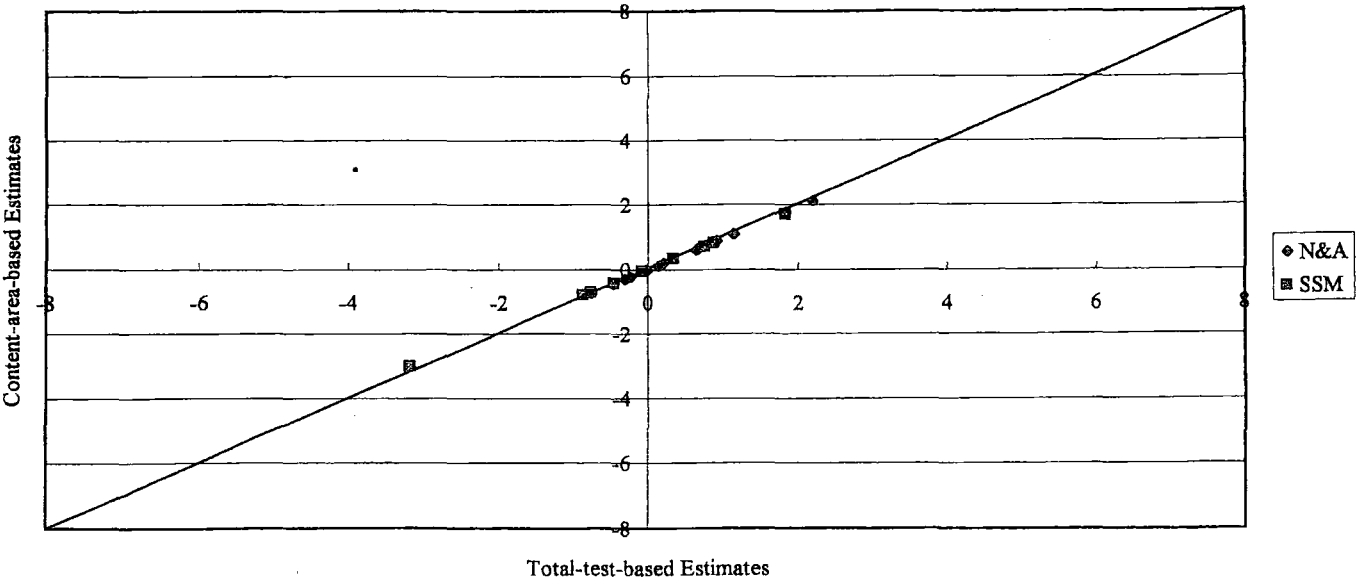
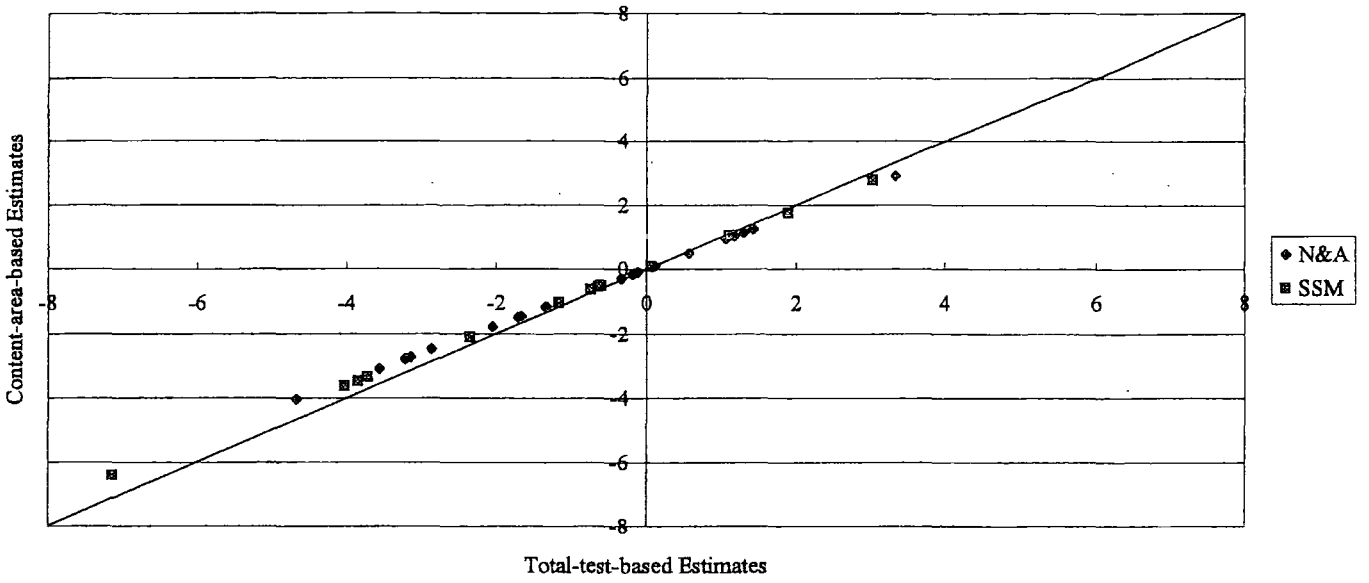


Figure 4.12
Plot of Content-area-based Difficulty Estimates vs.
Total-test-based Estimates for the High-ability Group
(a) Test-35 (r = 1.000)



(b) Test-24 (0.992)

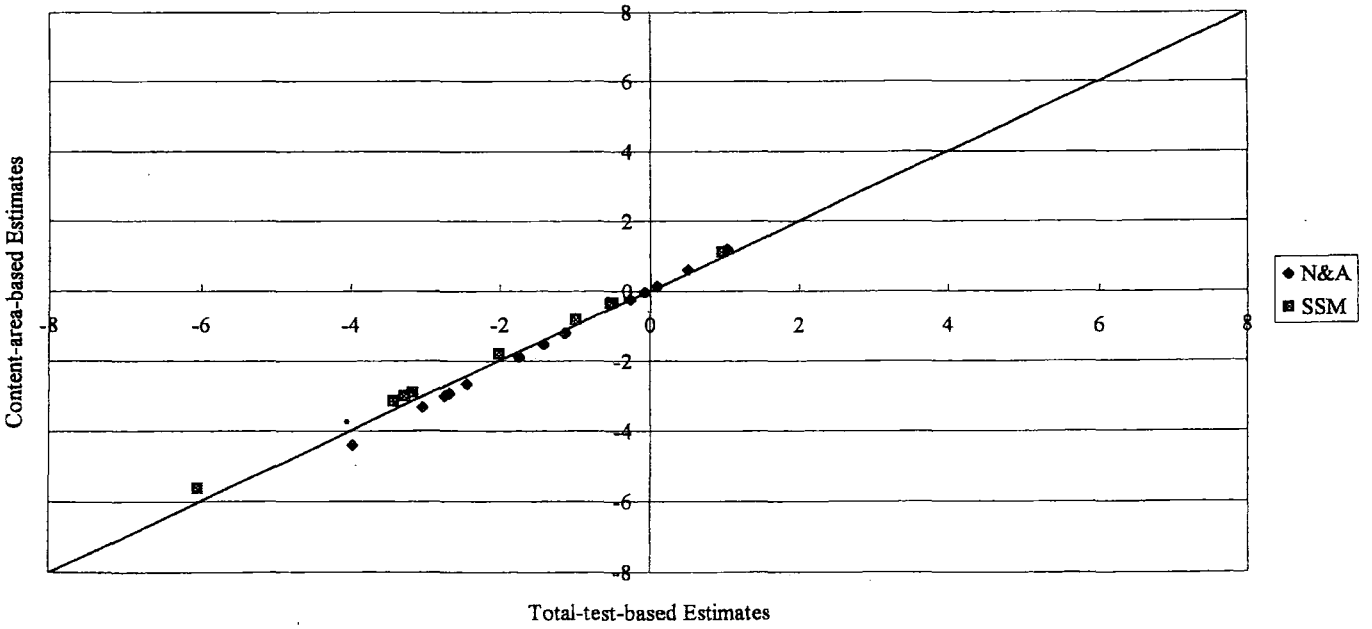
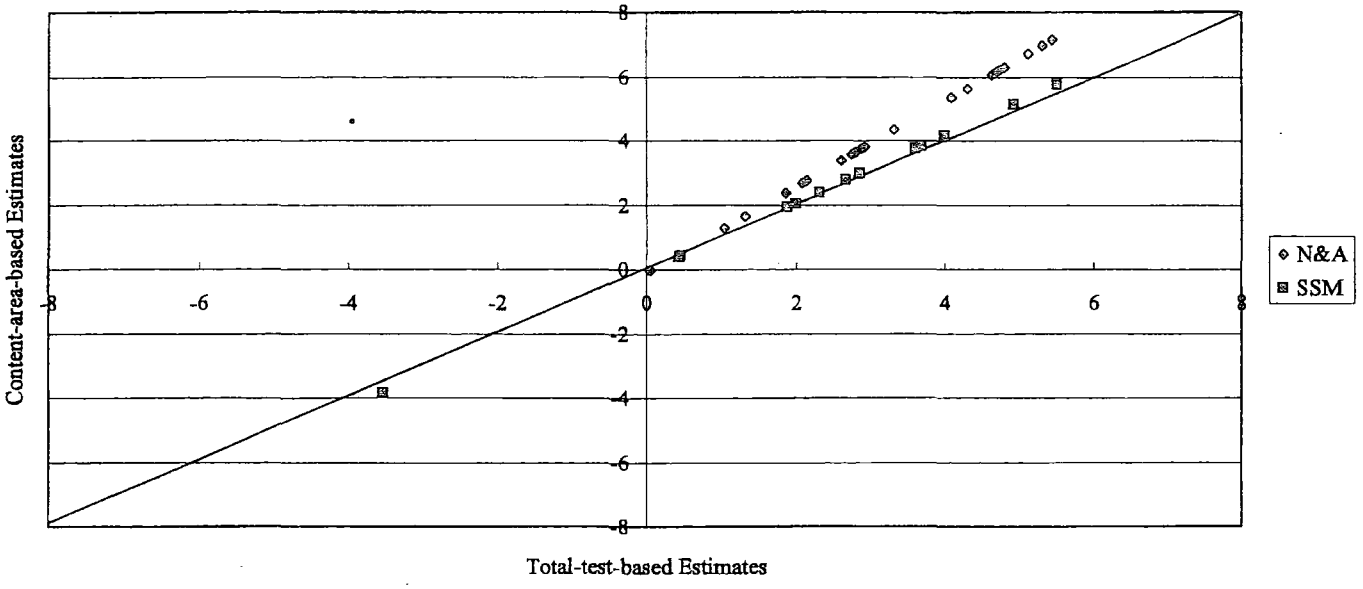


Figure 4.13
Plot of Content-area-based Difficulty Estimates vs.
Total-test-based Estimates for the Low-ability Group
(a) Test-35 (r = 0.982)



(b) Test-24 (r = 0.989)

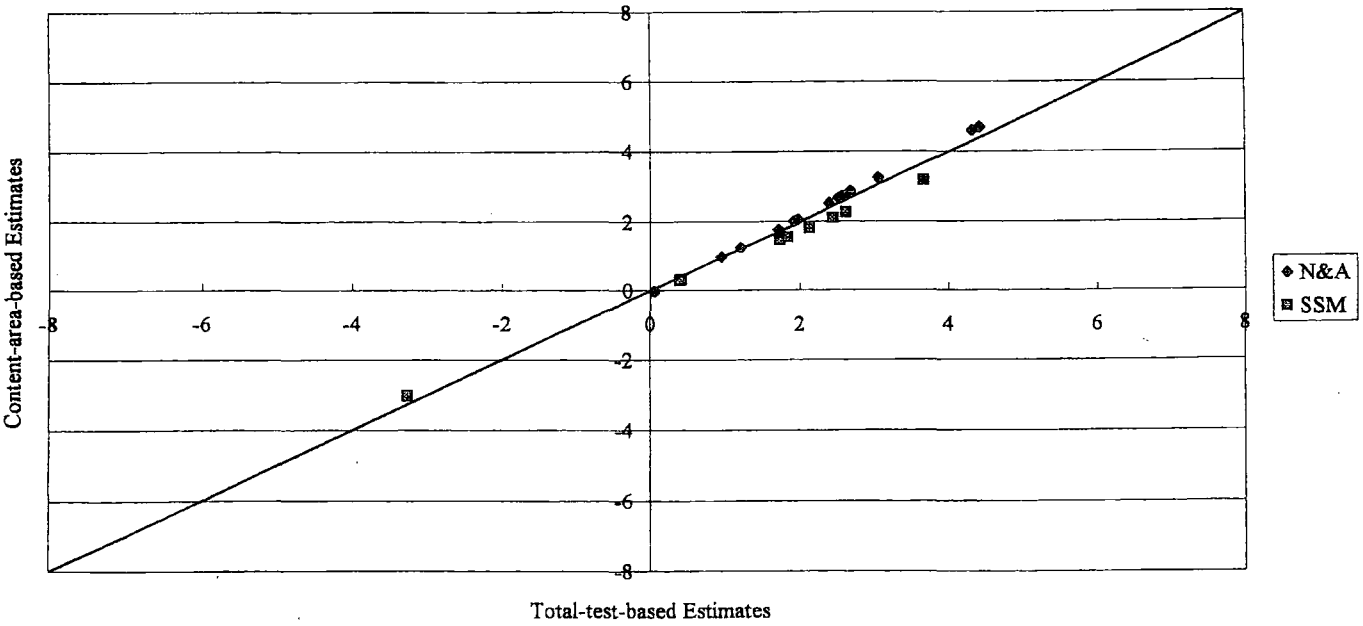
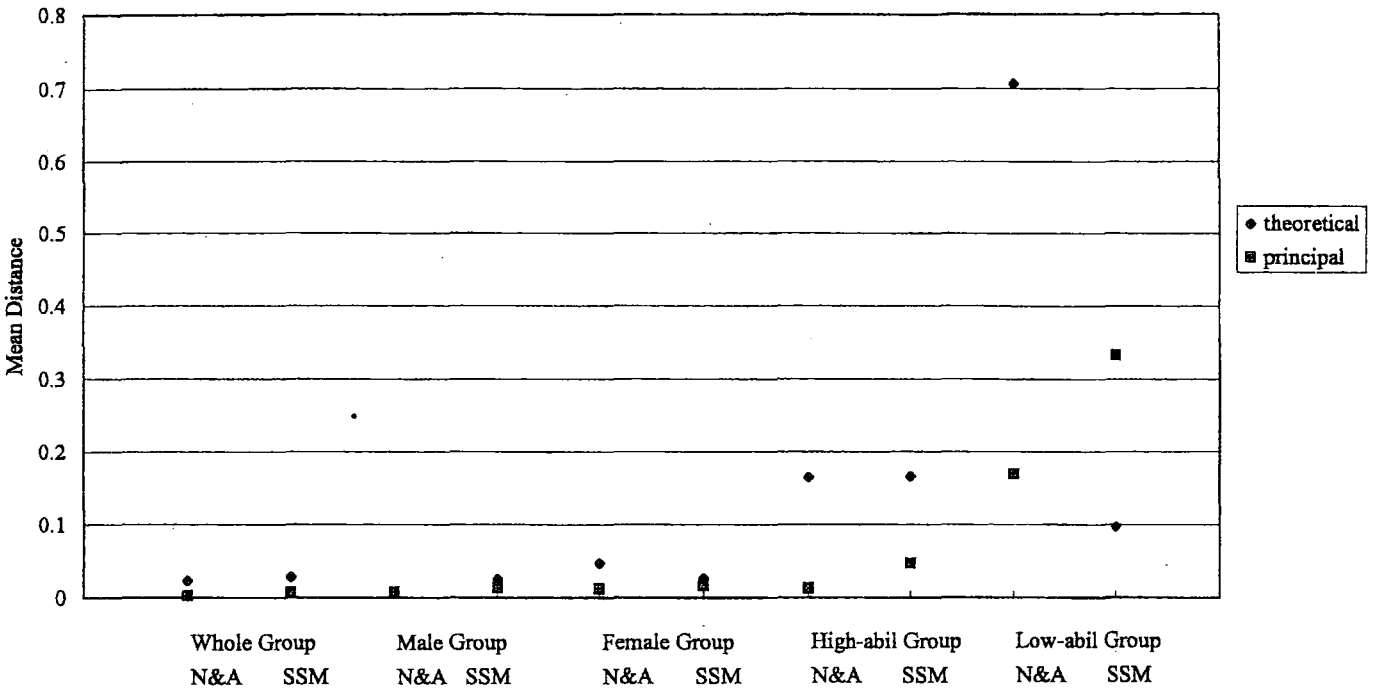


Figure 4.14
Mean Distance to the Principal and Theoretical Axes by Content Area
(a) Test-35



(b) Test-24

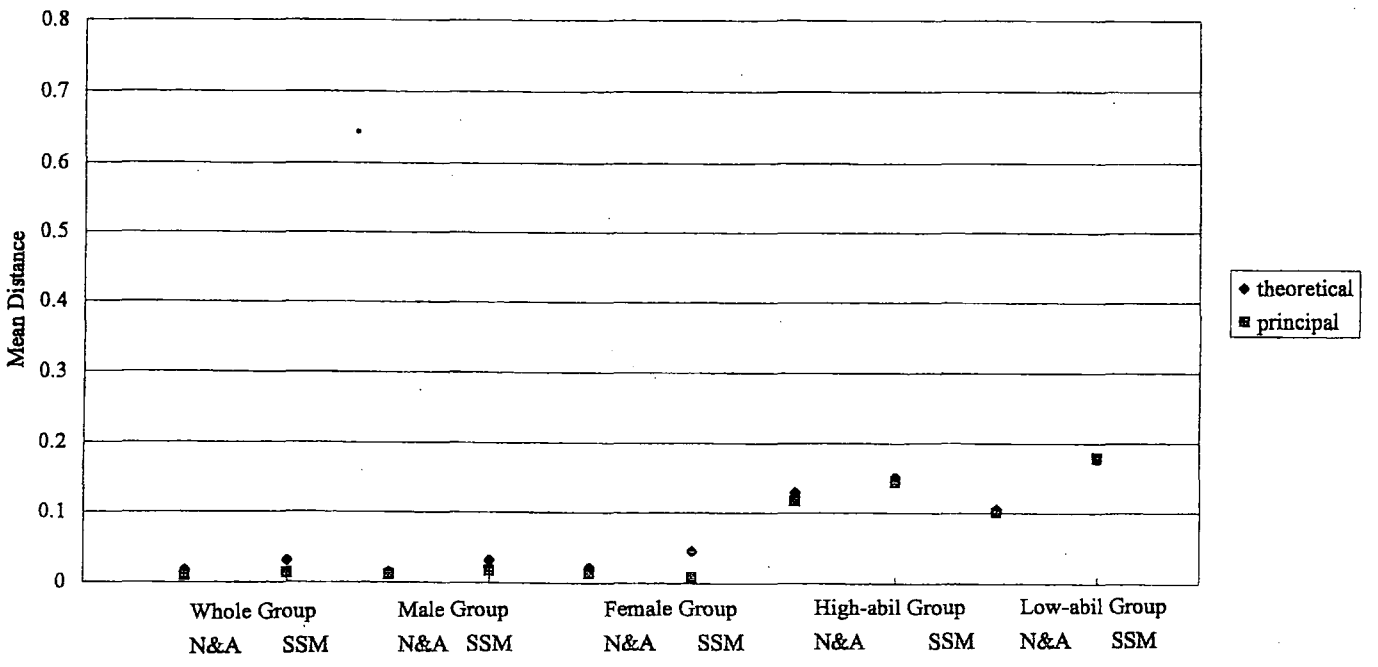
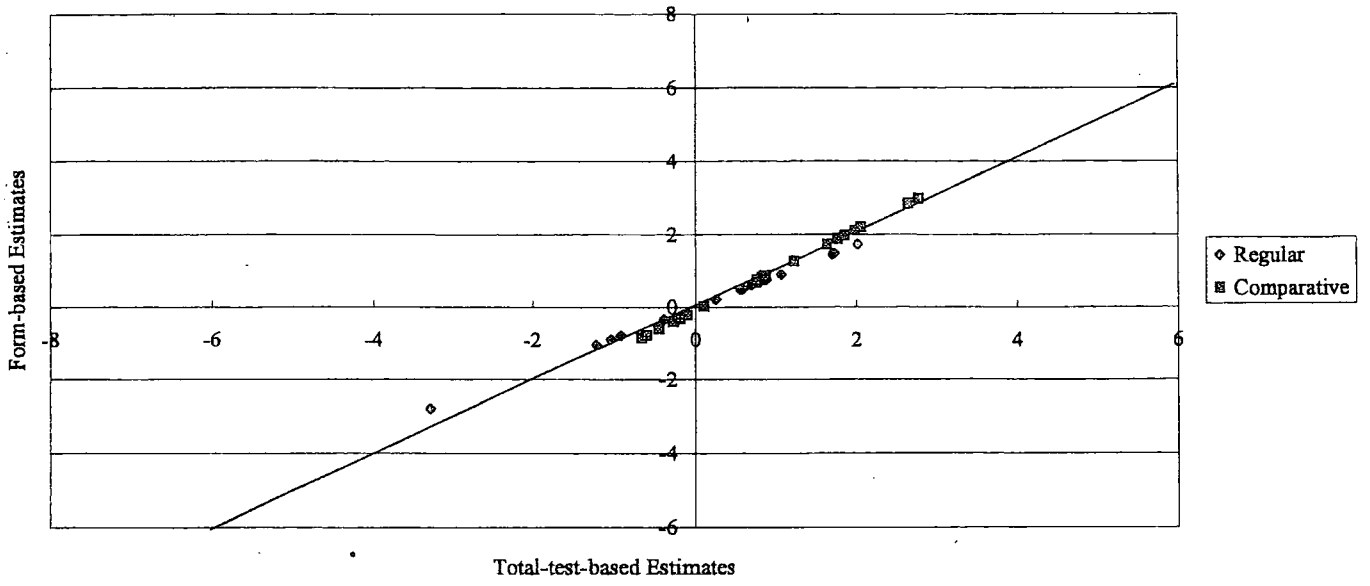


Figure 4.15
Plot of Form-based Difficulty Estimates vs.
Total-test-based Estimates for the Whole Group
(a) Test-35 (r = 0.992)



(b) Test-24 (r = 1.000)

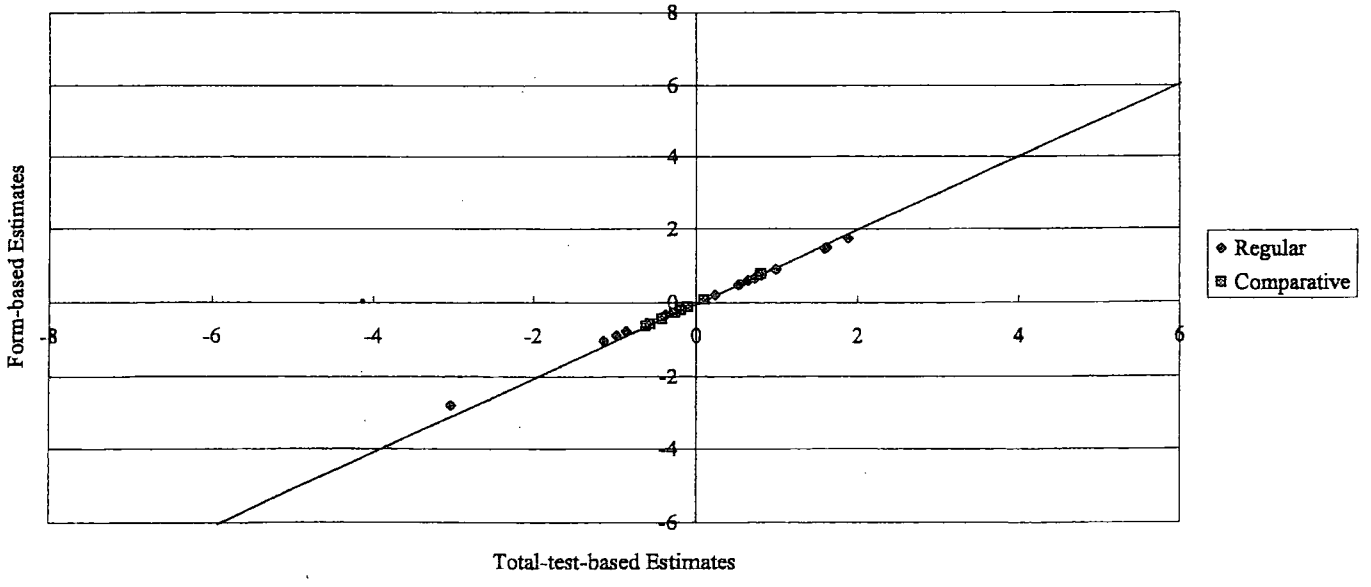
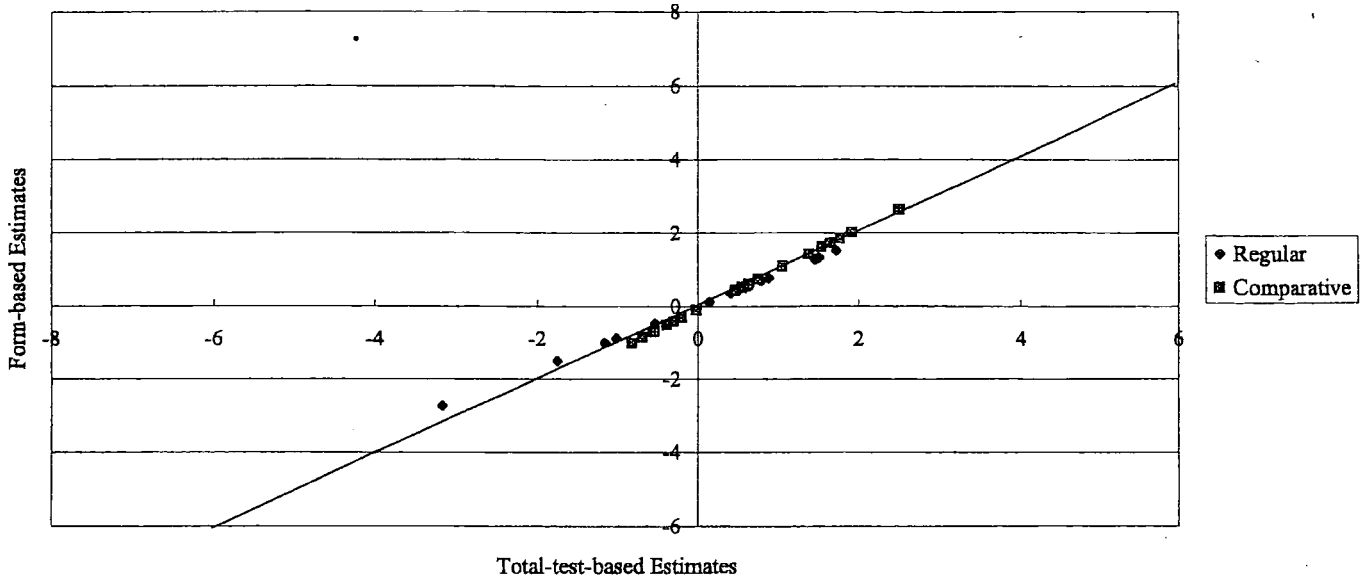
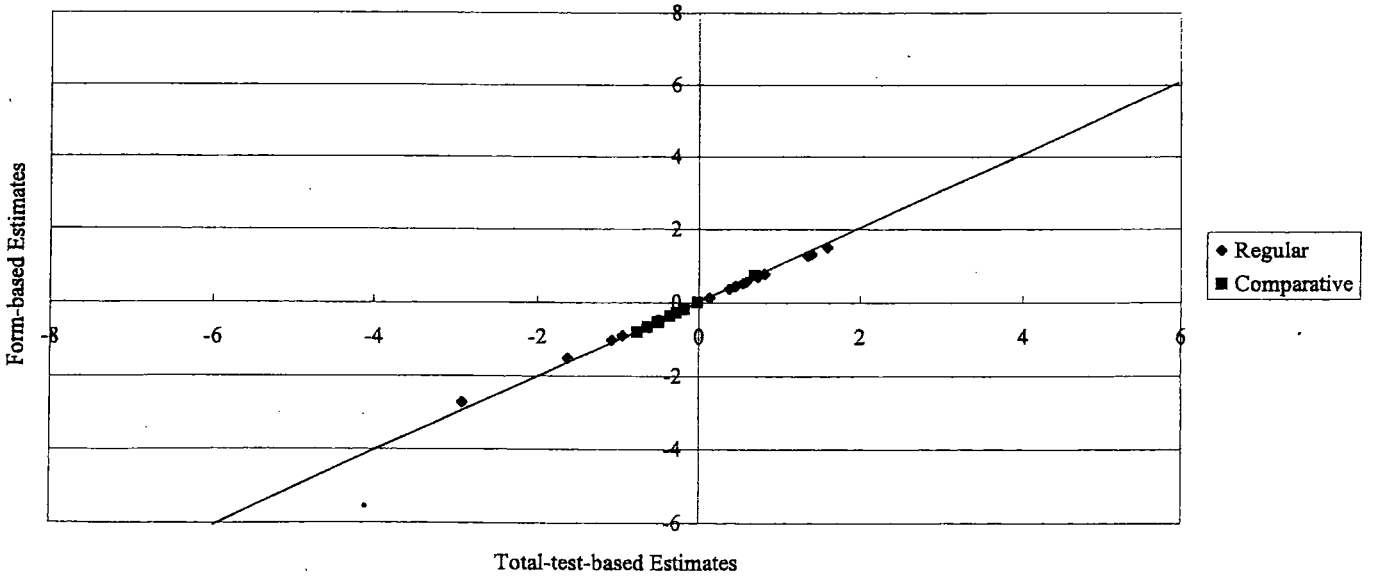


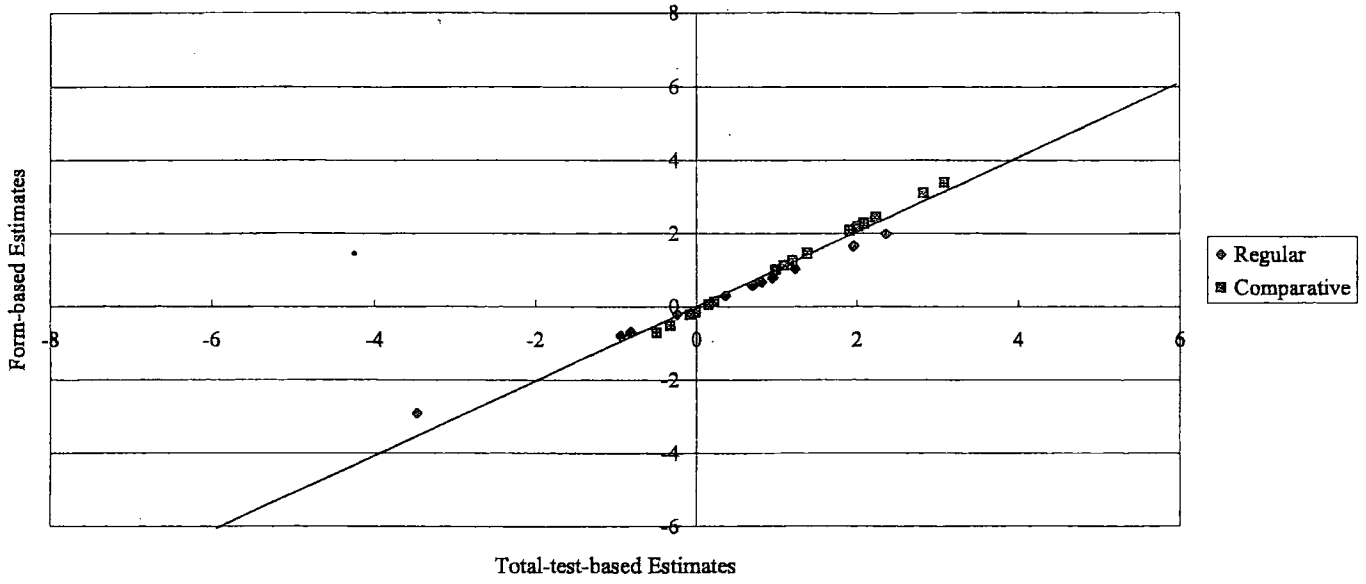
Figure 4.16
Plot of Form-based Difficulty Estimates vs.
Total-test-based Estimates for the Male Group
(a) Test-35 (r = 0.993)



(b) Test-24 (r = 1.000)



Firure 4.17
Plot of Form-based Difficulty Estimates vs.
Total-test-based Estimates for the Female Group
(a) Test-35 (r = 0.988)



(b) Test-24 (r = 0.999)

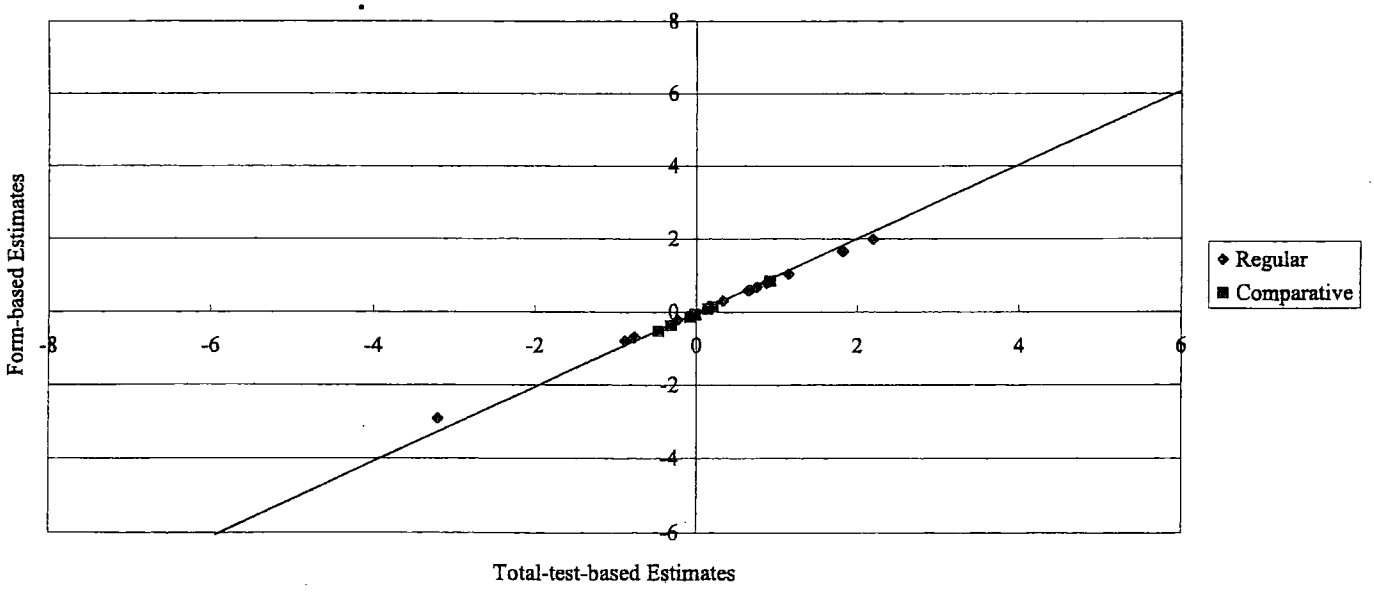
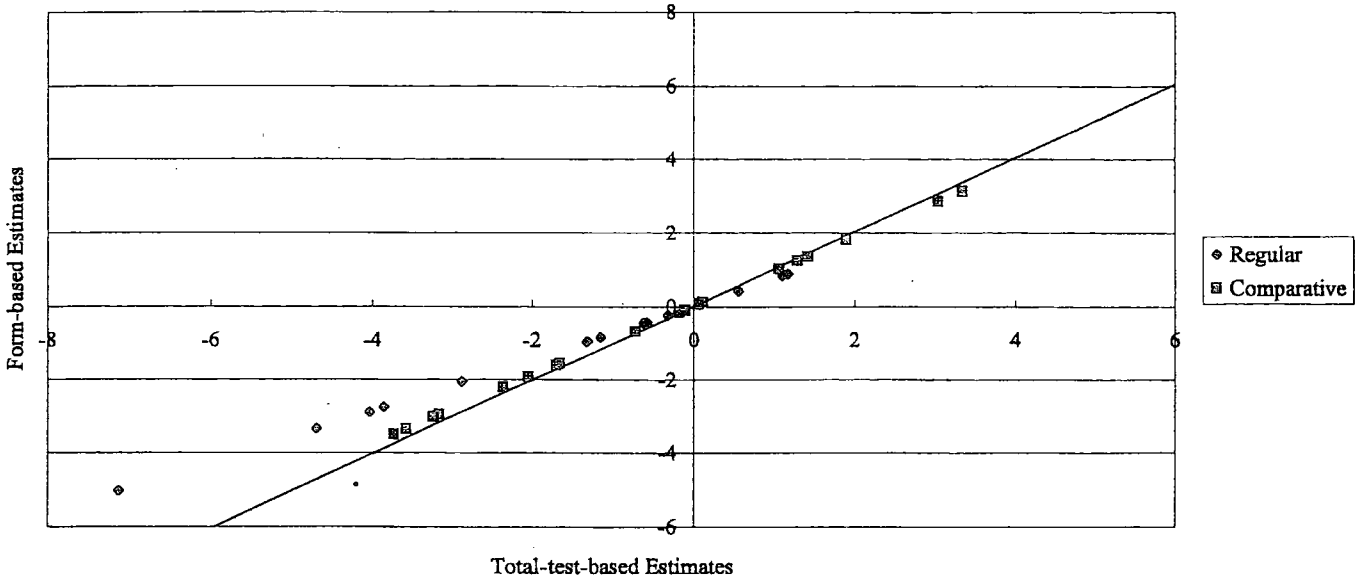


Figure 4.18
Plot of Form-based Difficulty Estimates vs.
Total-test-based Estimates for the High-ability Group
(a) Test-35 (r = 0.989)



(b) Test-24 (r = 0.971)

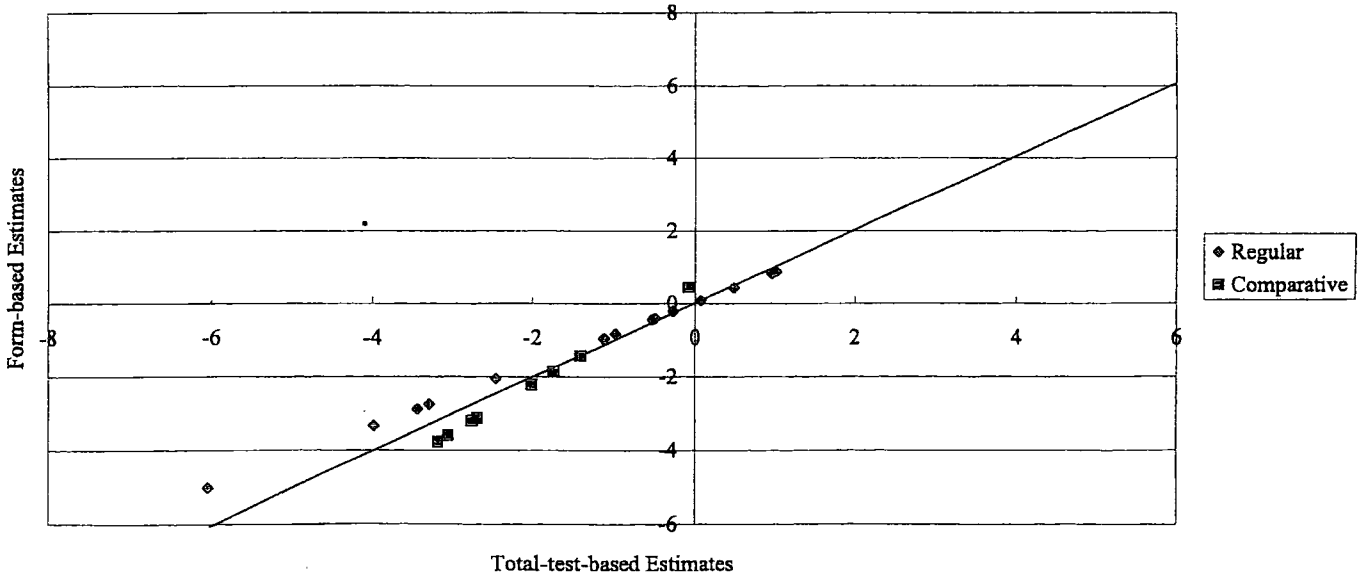
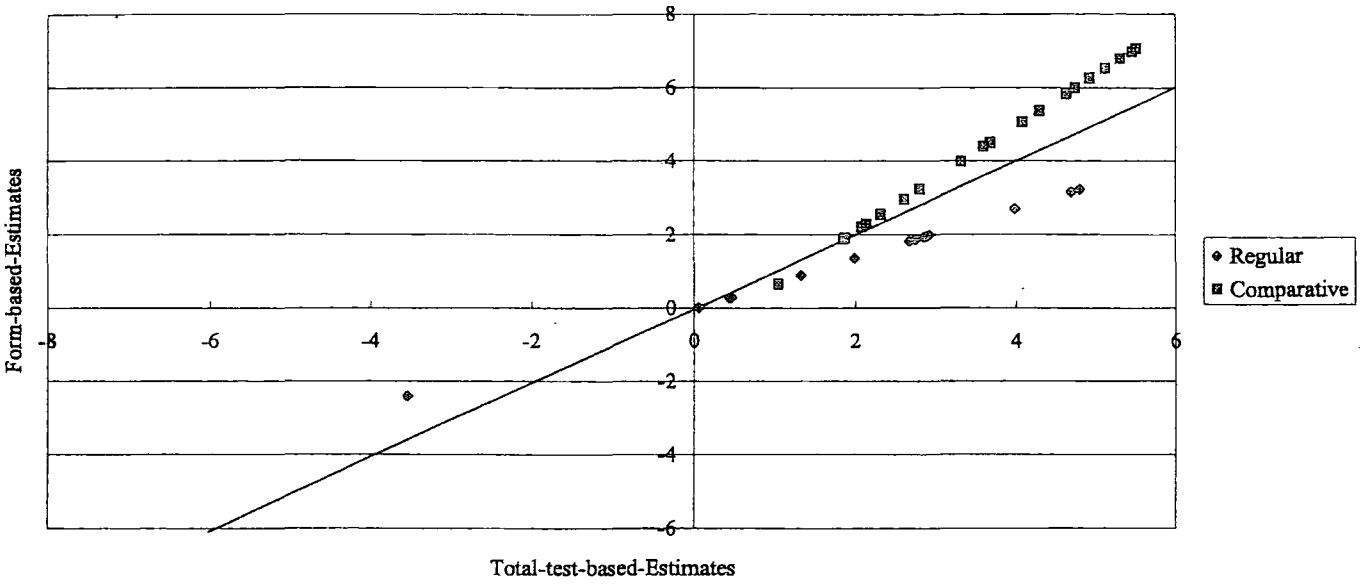


Figure 4.19
Plot of Form-based Difficulty Estimates vs.
Total-test-based Estimates for the Low-ability Group
(a) Test-35 (r = 0.915)



(b) Test-24 (r = 0.996)

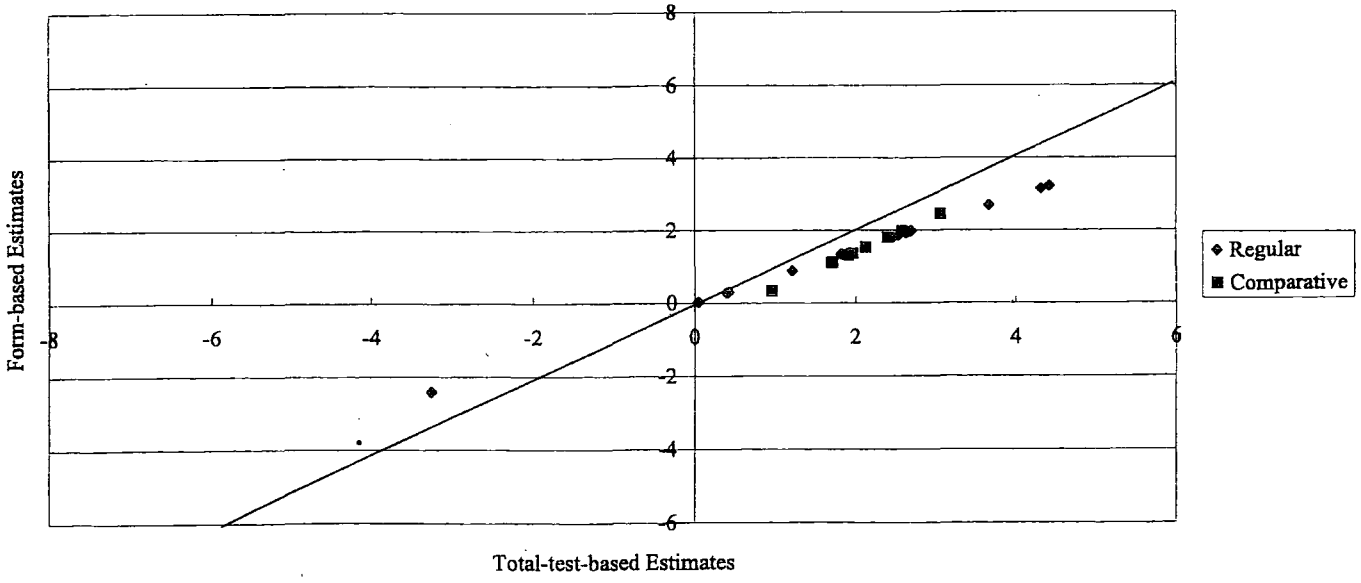
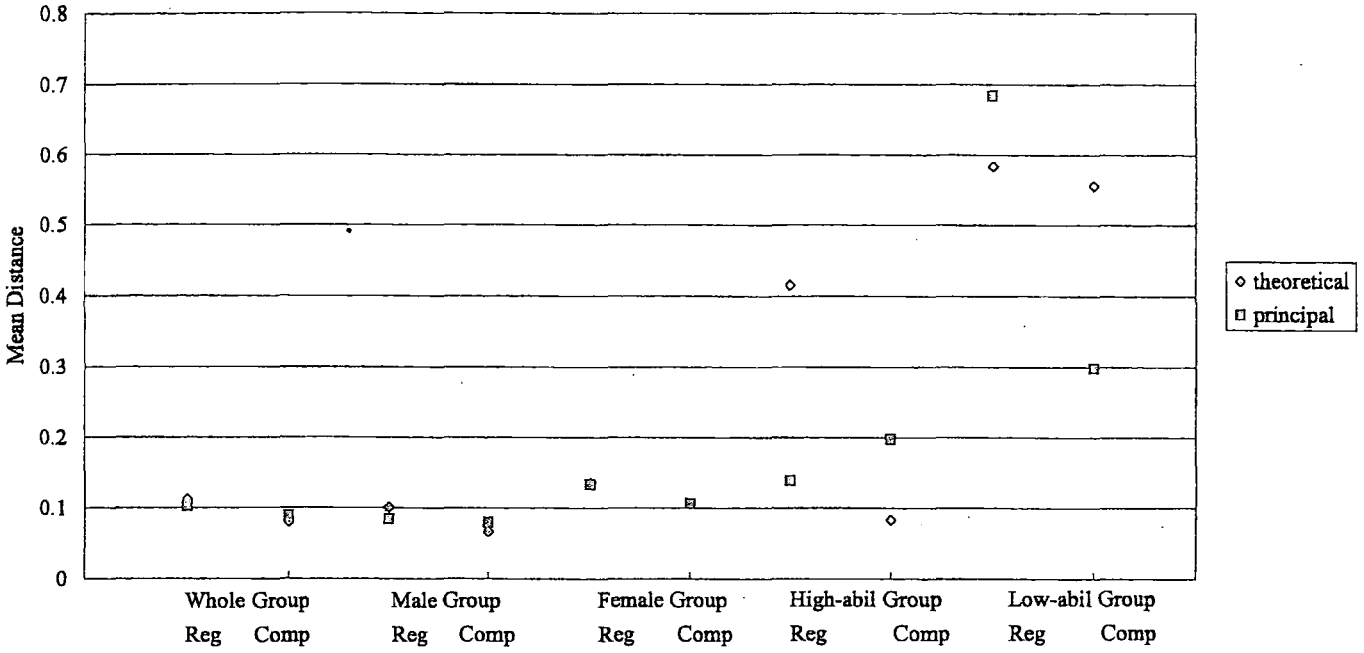


Figure 4.20
Mean Distances to the Principal and Theoretical Axes by Item Form
(a) Test-35



(b) Test-24

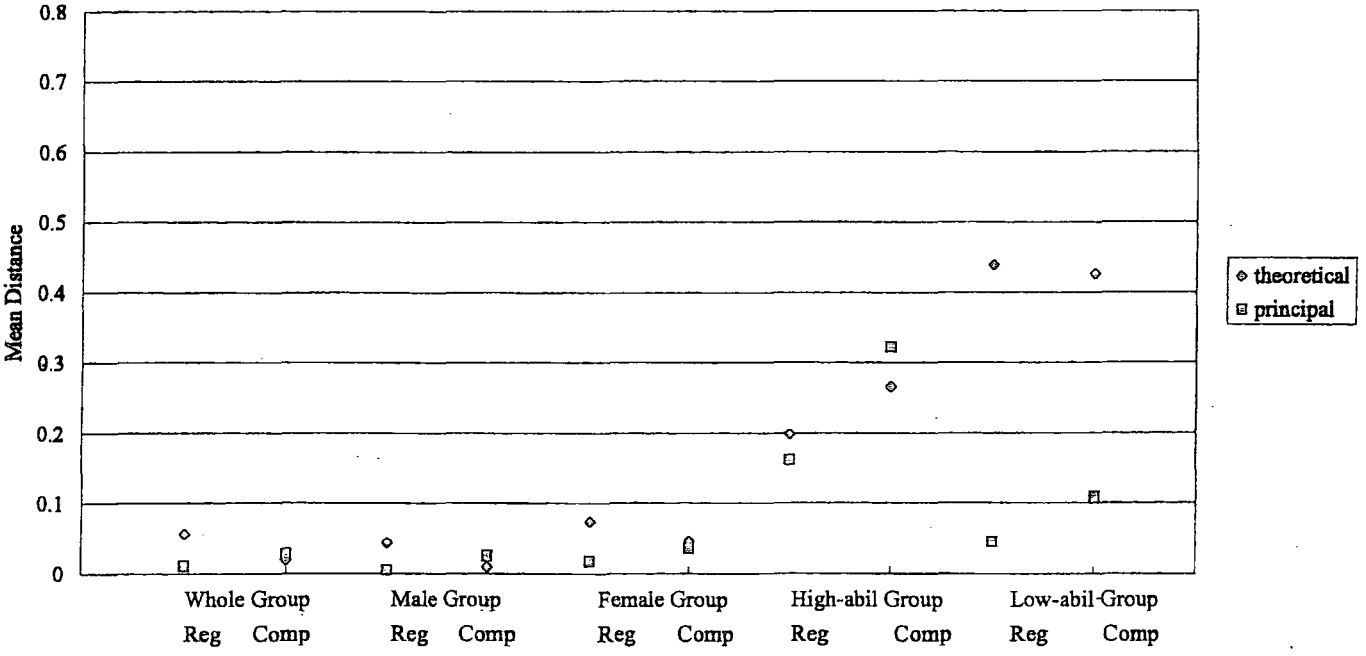
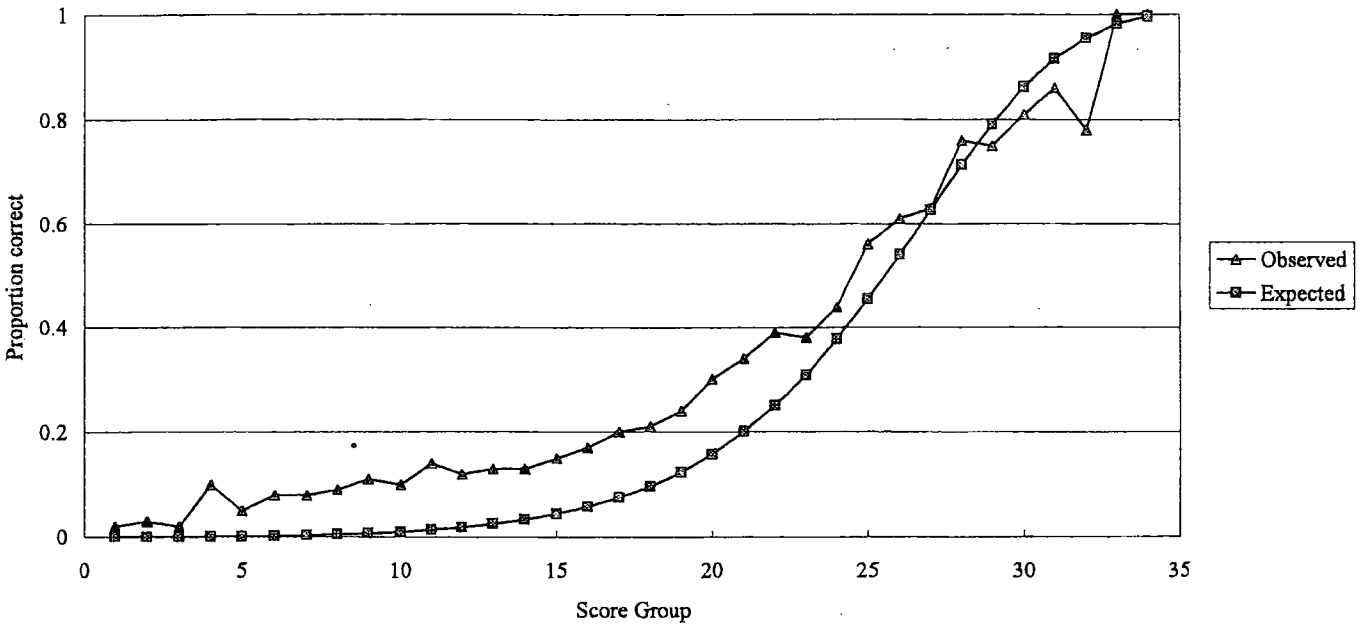
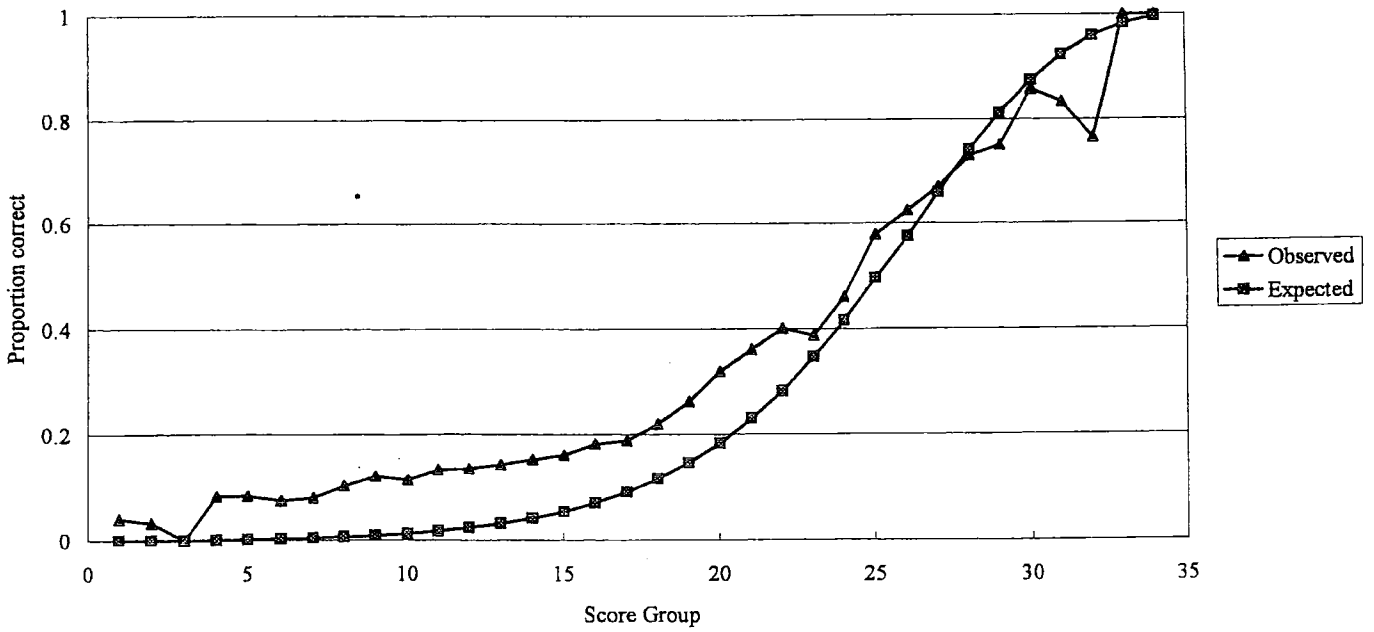


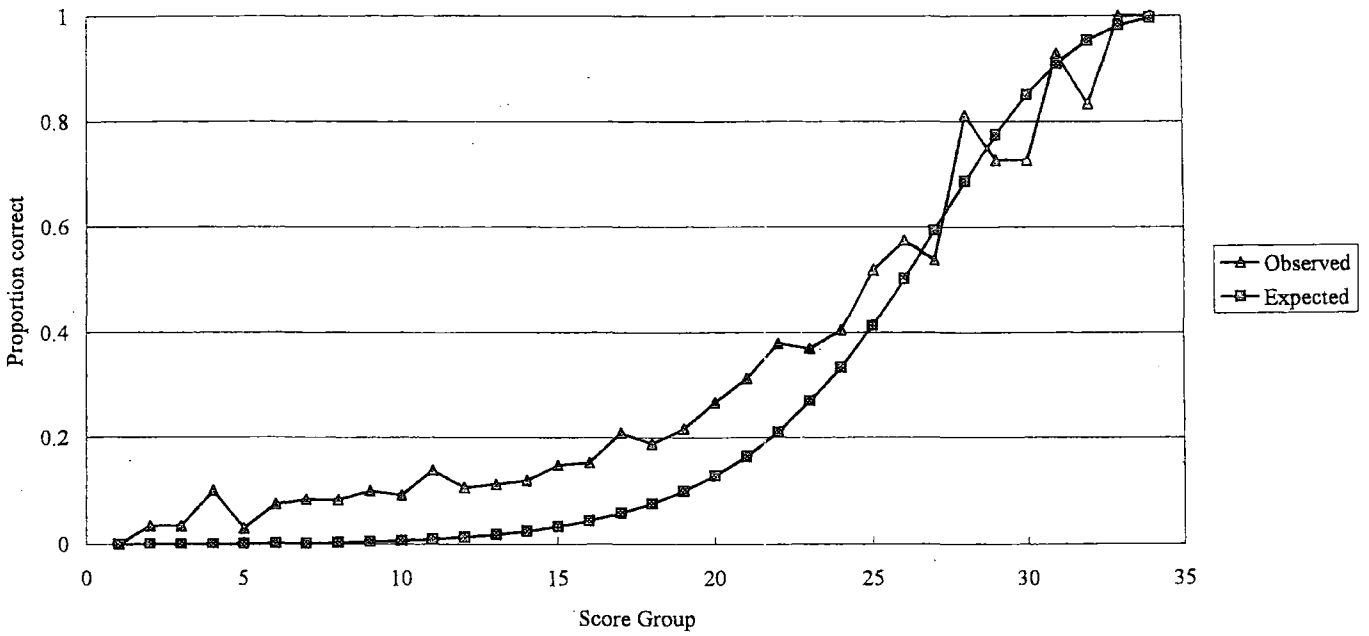
Figure 4.21
ICCs for Item 13 of Test-35
(a) Whole Group (b = 2.015)



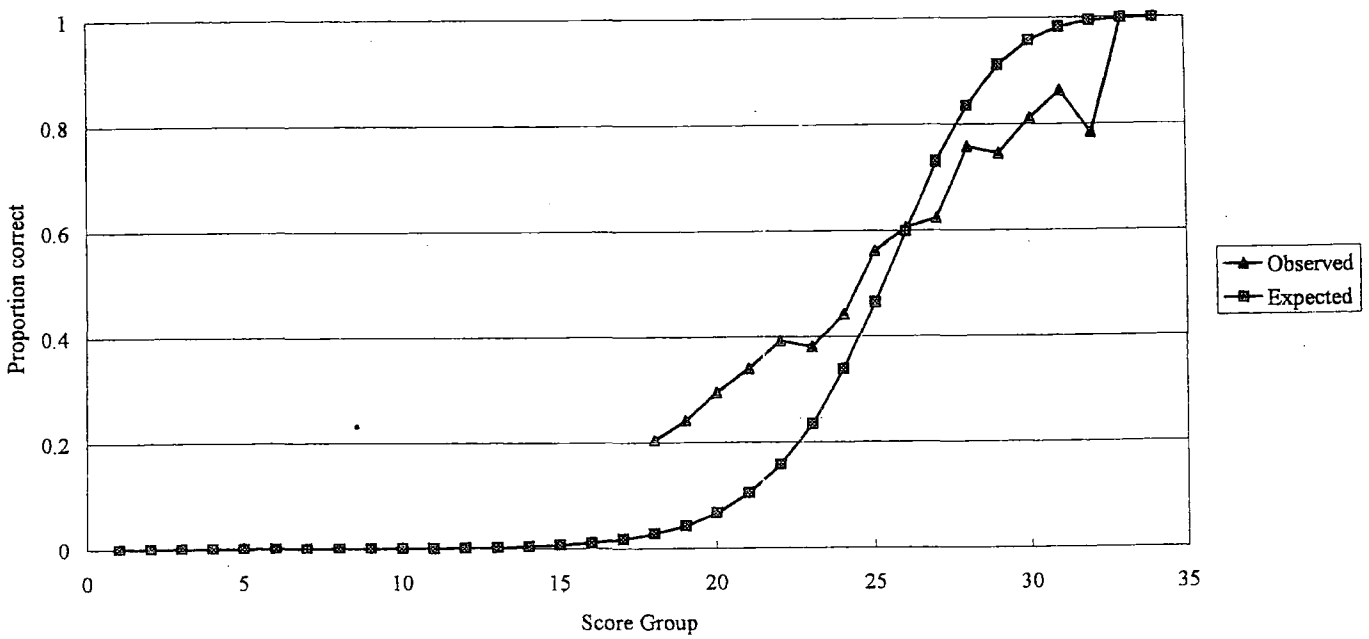
(b) Male Group (b = 1.720)



(c) Female Group (b = 2.355)



(d) High-ability Group (b = 1.173)



(e) Low-ability Group (b = 4.796)

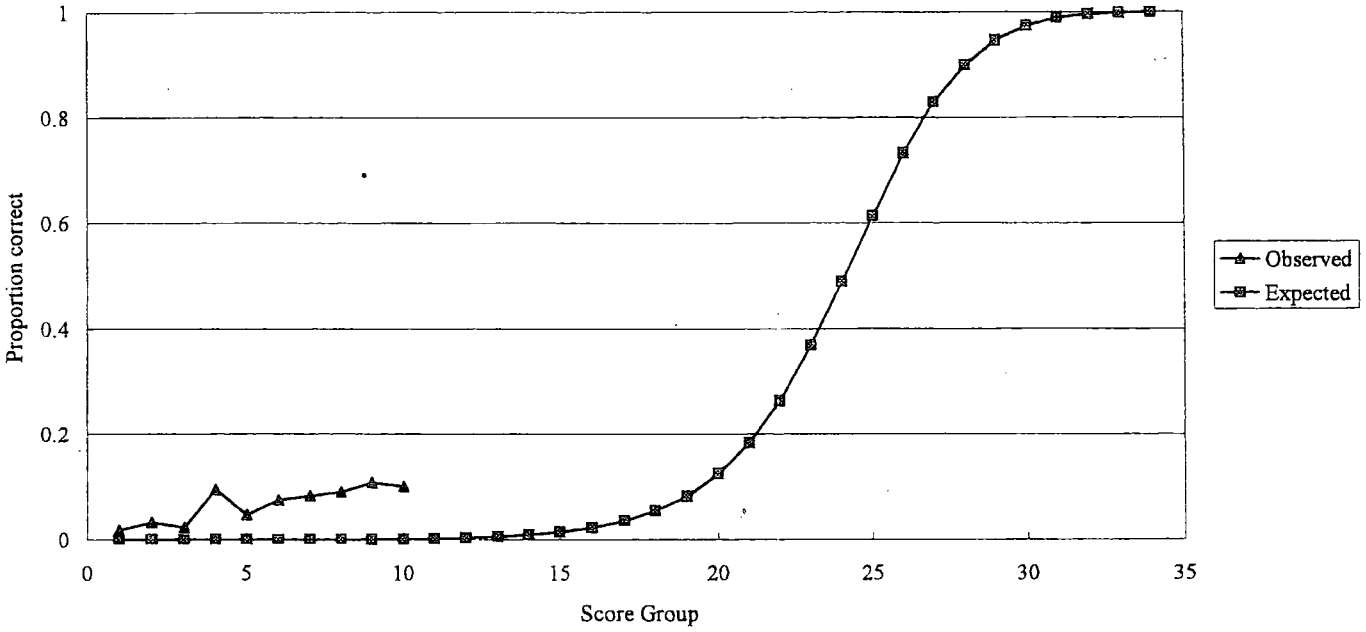
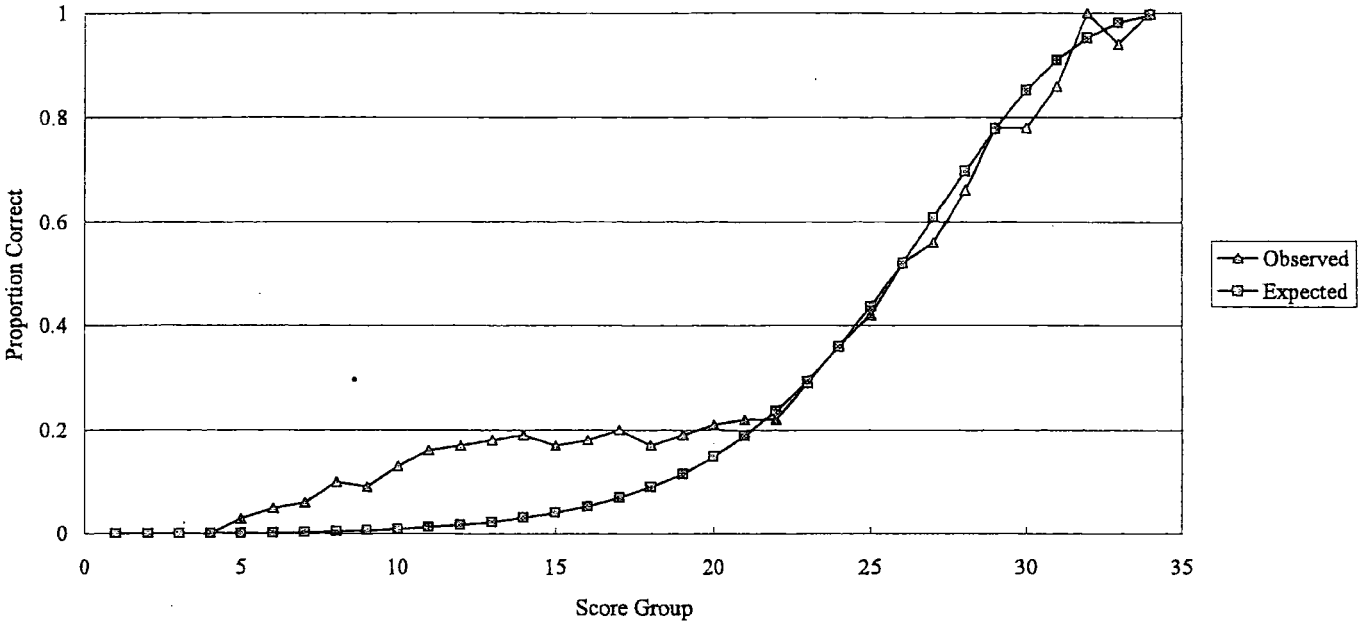
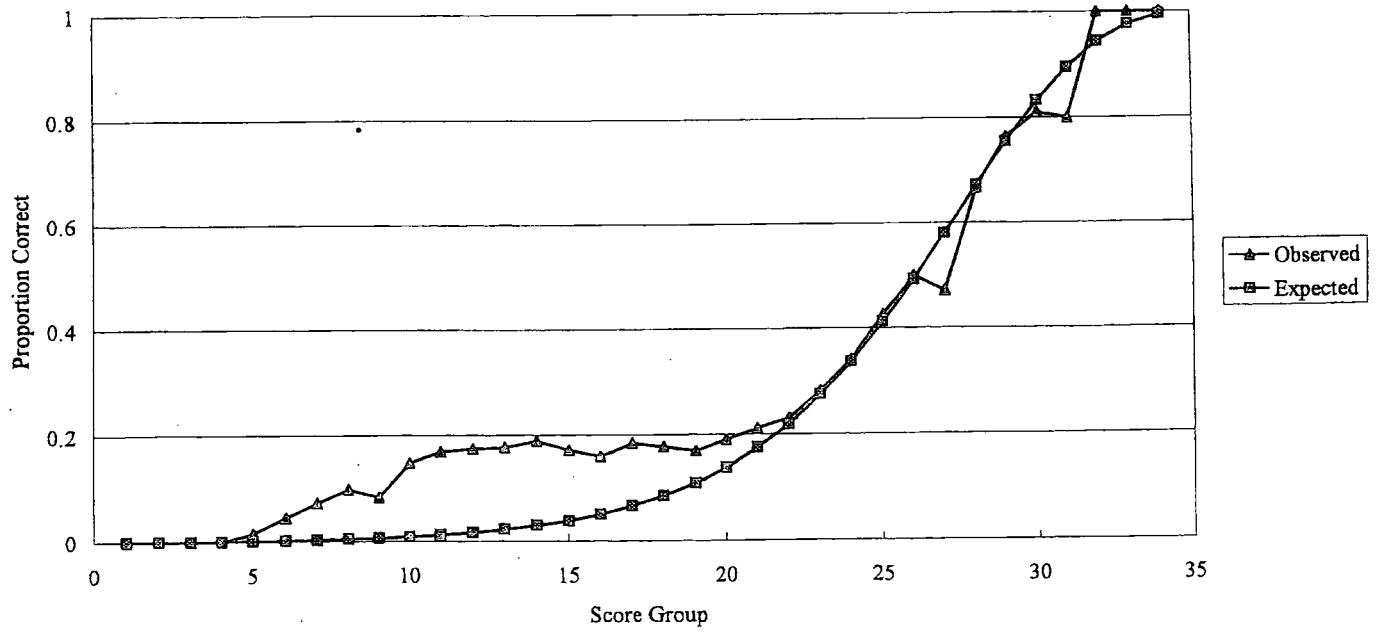


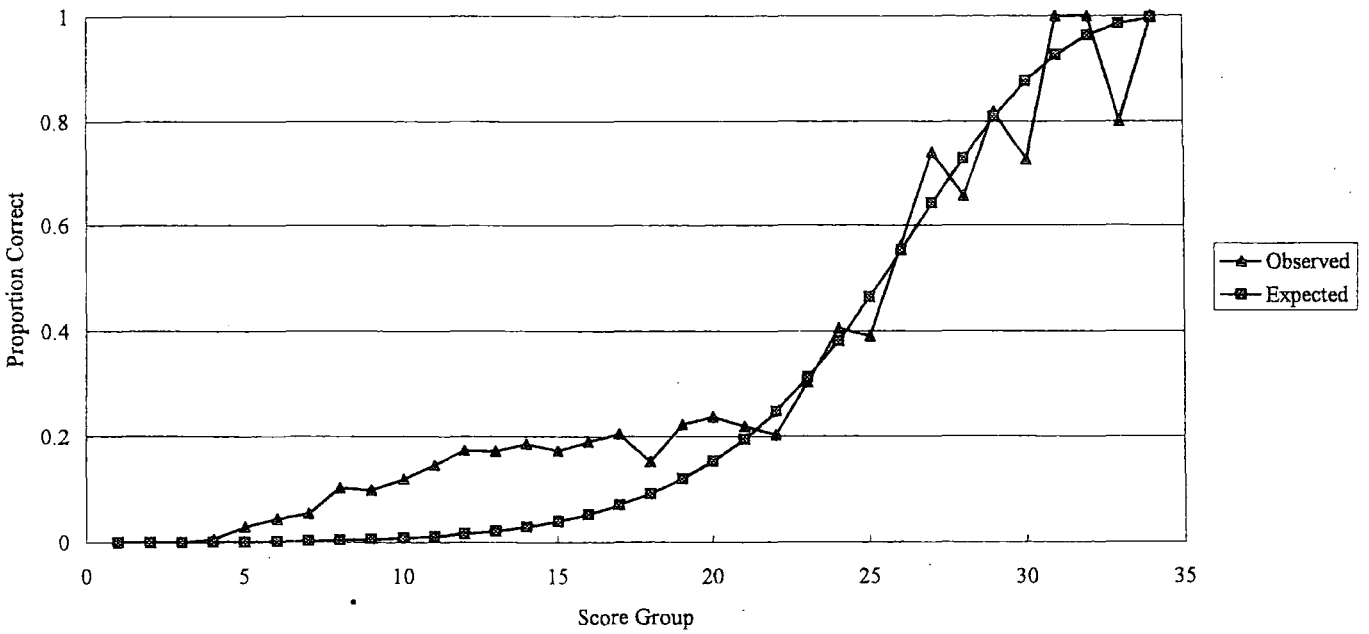
Figure 4.22
ICCs for Item 33 of Test-35
(a) Whole Group (b = 2.060)



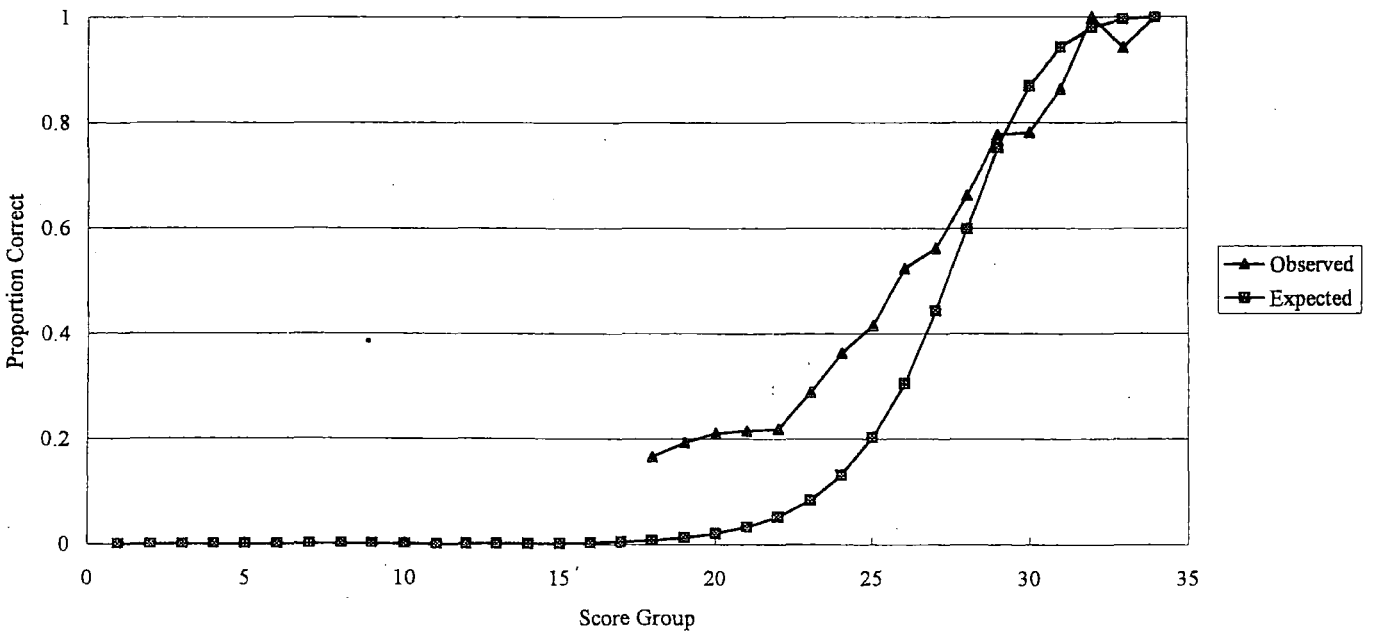
(b) Male Group (b = 1.918)



(c) Female Group (b = 2.235)



(d) High-ability Group (b = 1.894)



(e) Low-ability Group (b = 4.919)

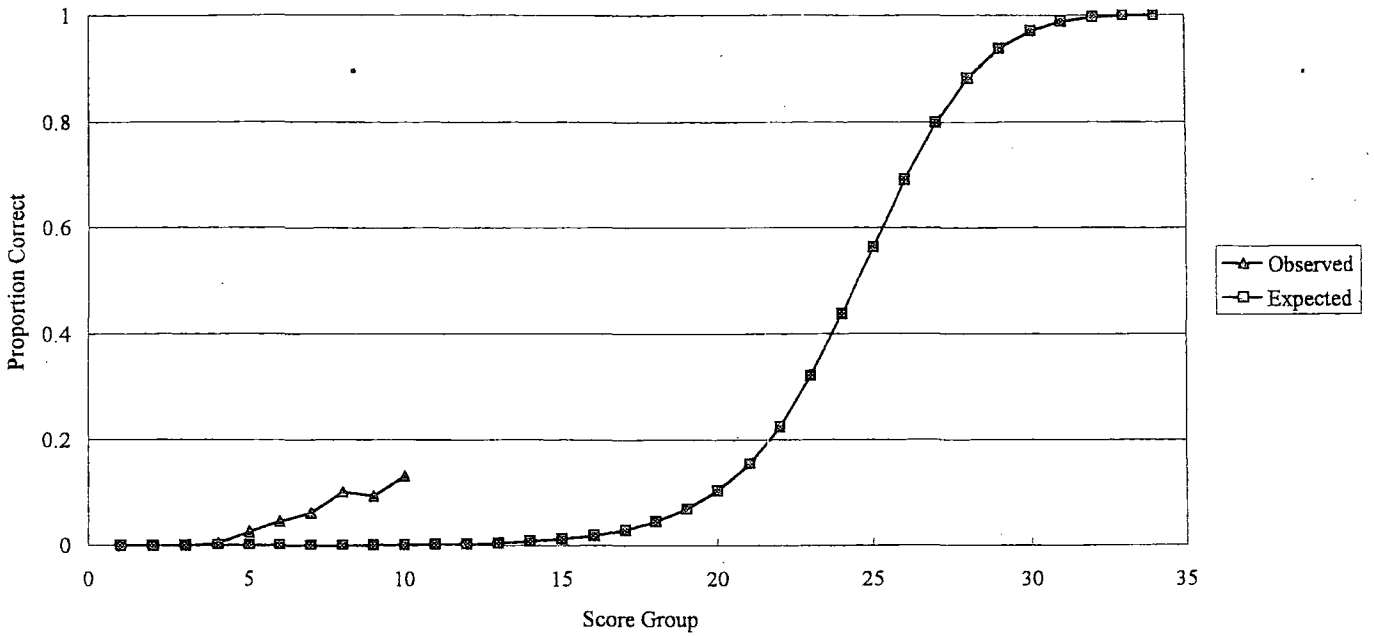
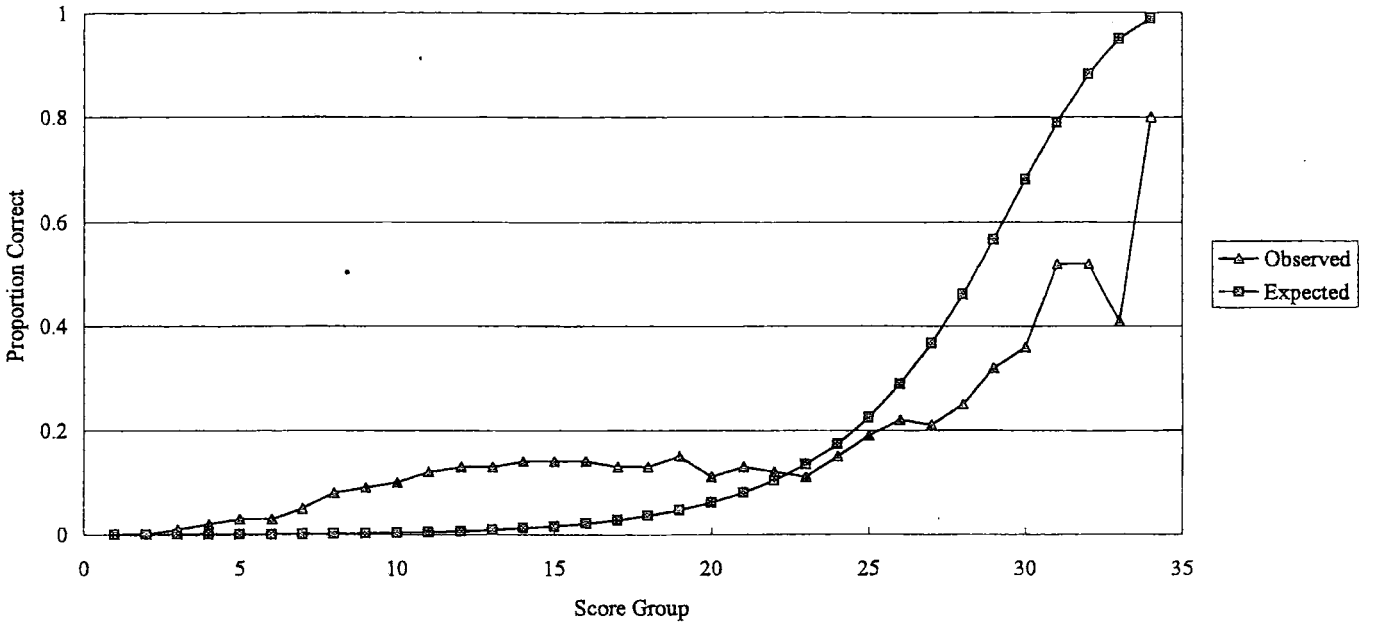
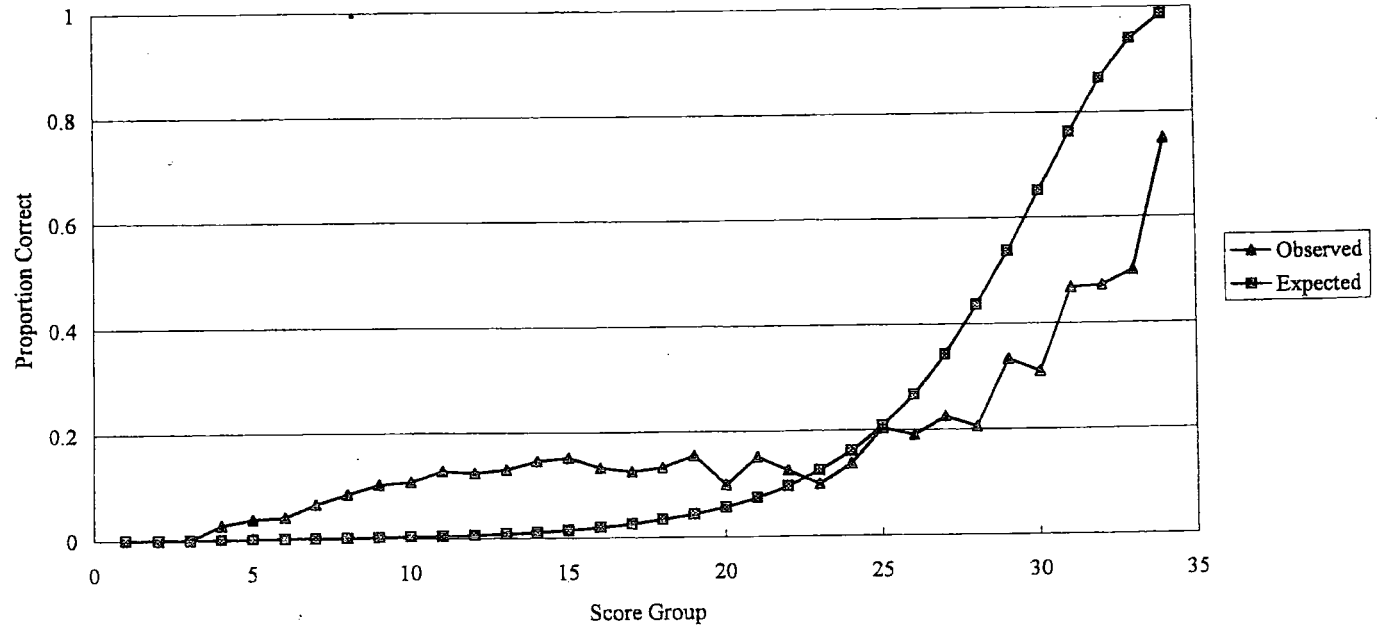


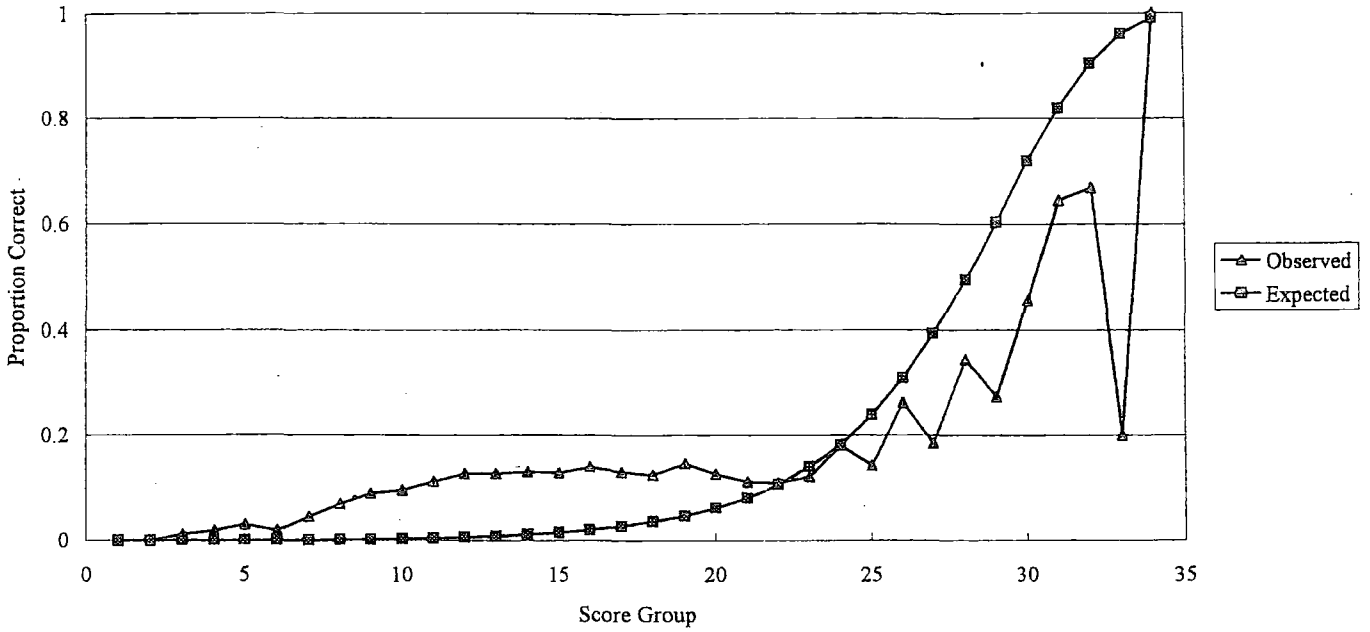
Figure 4.23
ICCs for Item 34 of Test-35
(a) Whole Group (b = 2.640)



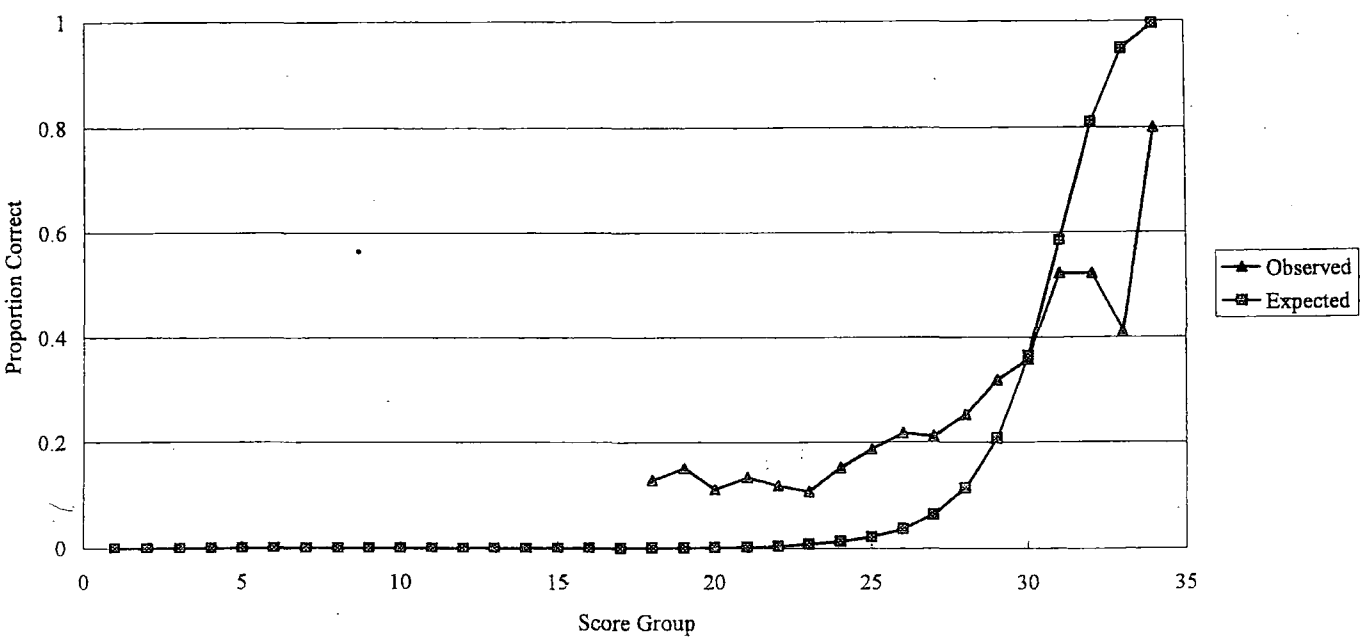
(b) Male Group (b = 2.492)



(c) Female Group (b = 2.833)



(d) High-ability Group (b = 3.336)



(e) Low-ability Group (b = 5.303)

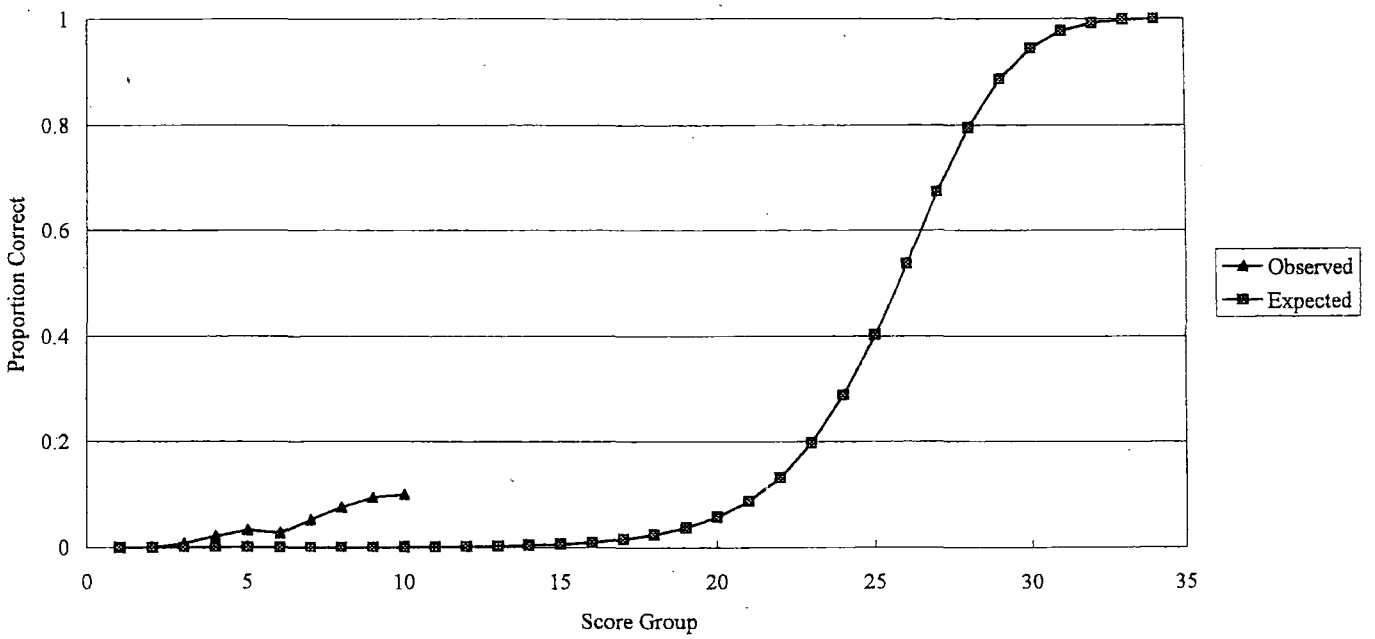
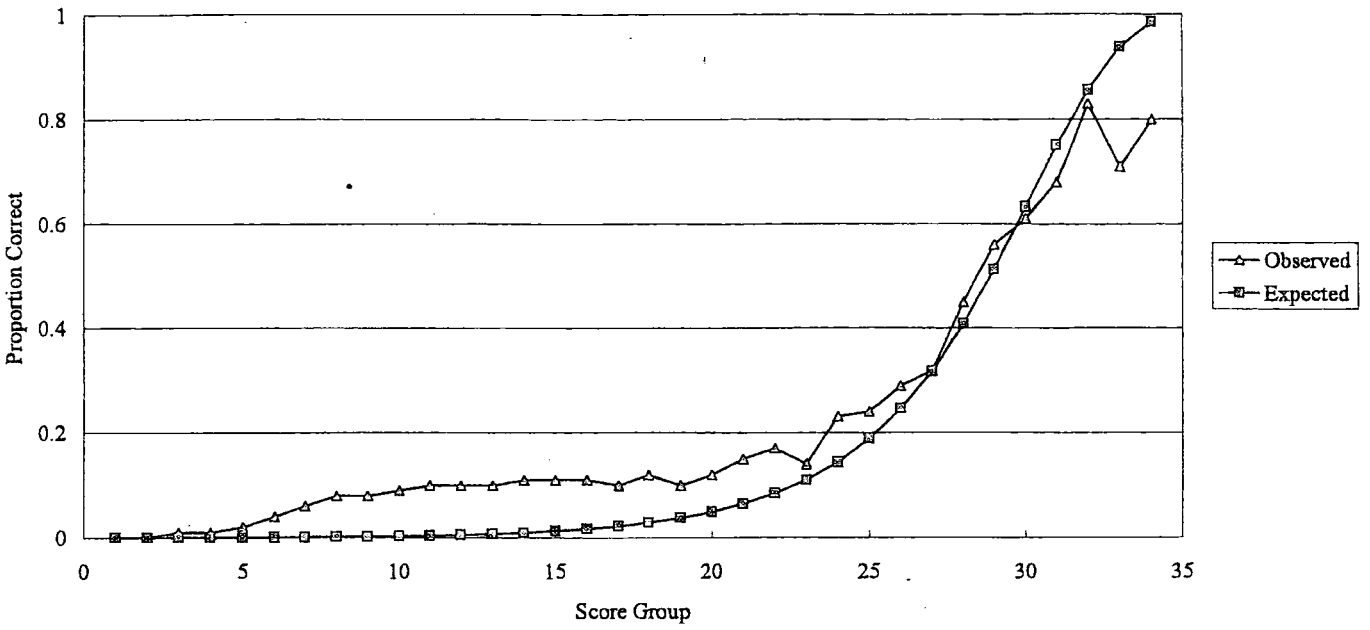
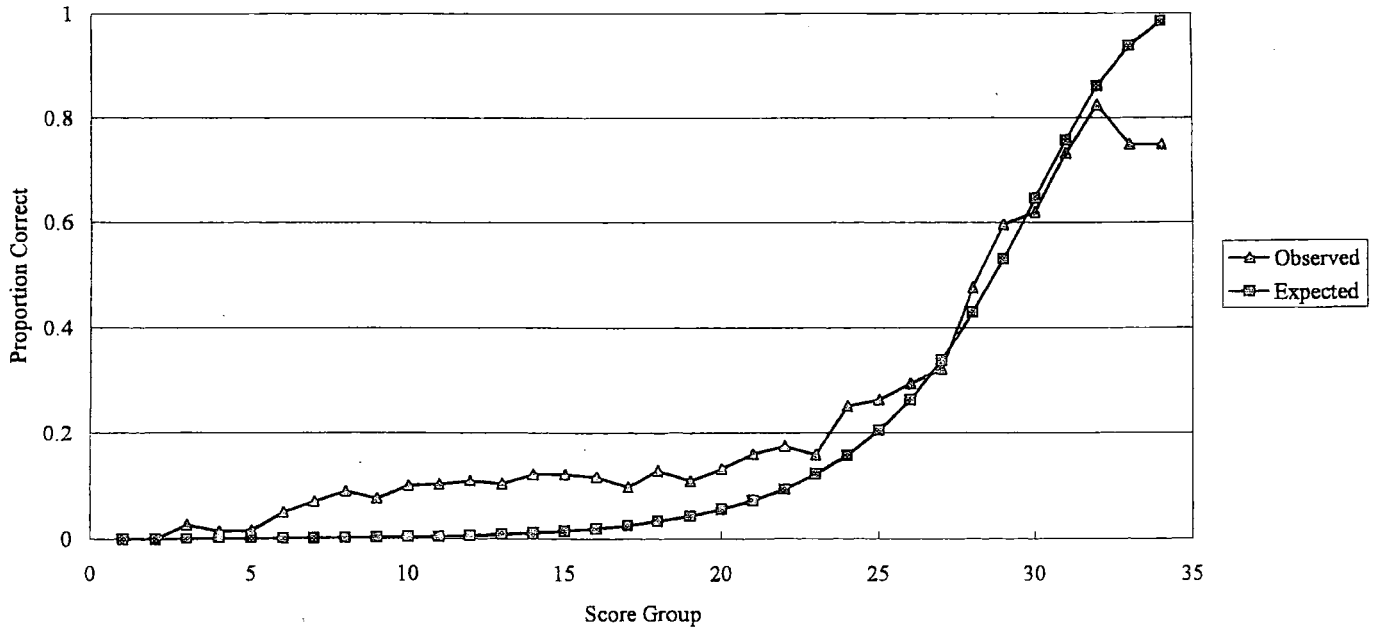


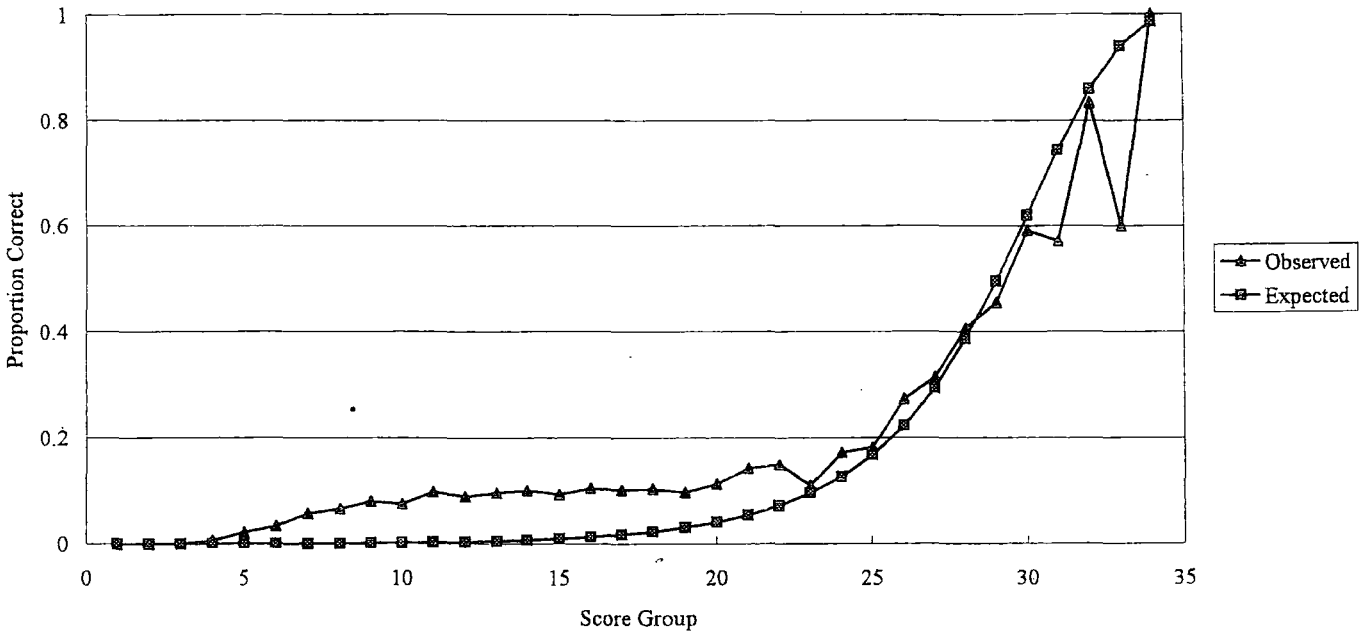
Figure 4.24
ICCs for Item 35 of Test-35
(a) Whole Group (b = 2.769)



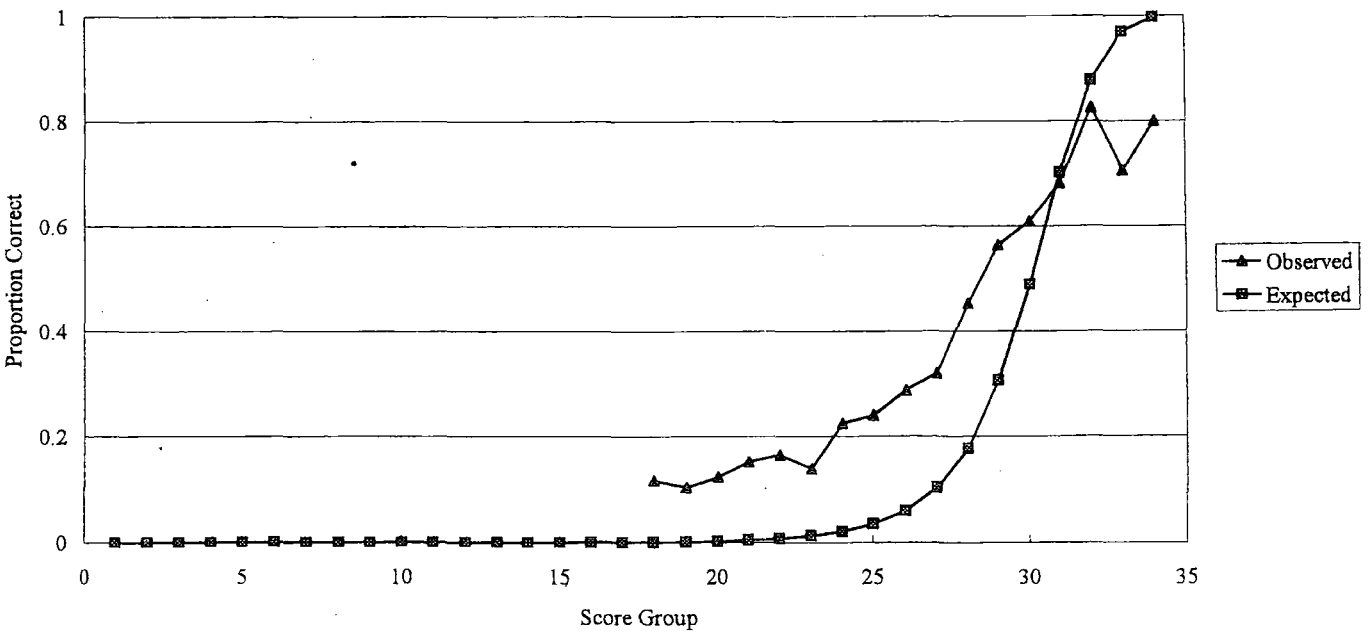
(b) Male Group (b = 2.506)



(c) Female Group (b = 3.091)



(d) High-ability Group (b = 3.034)



(e) Low-ability Group ($\hat{b}$ = 5.500)

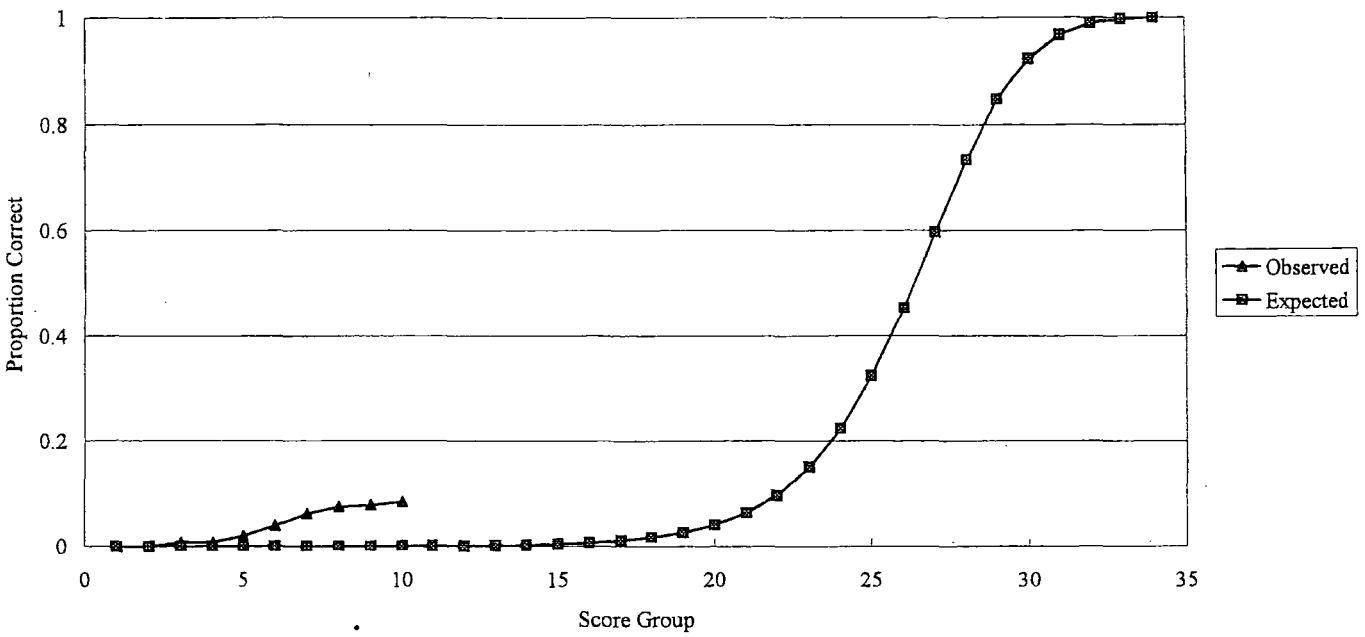
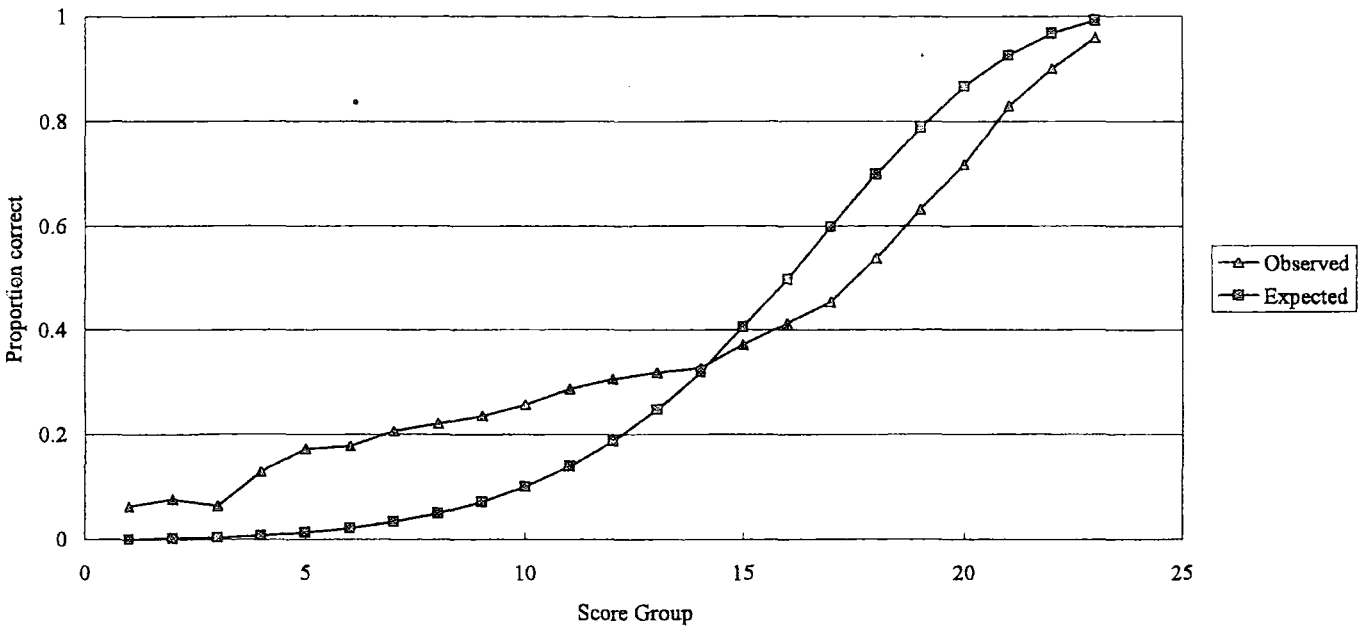
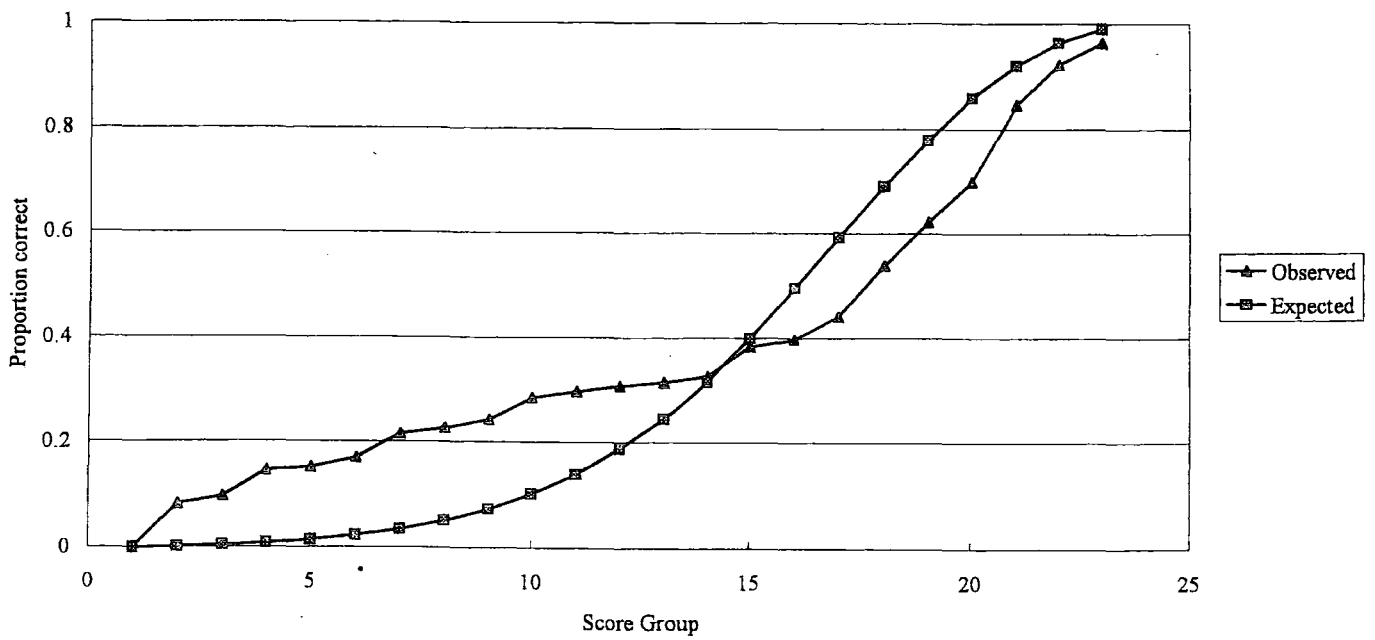


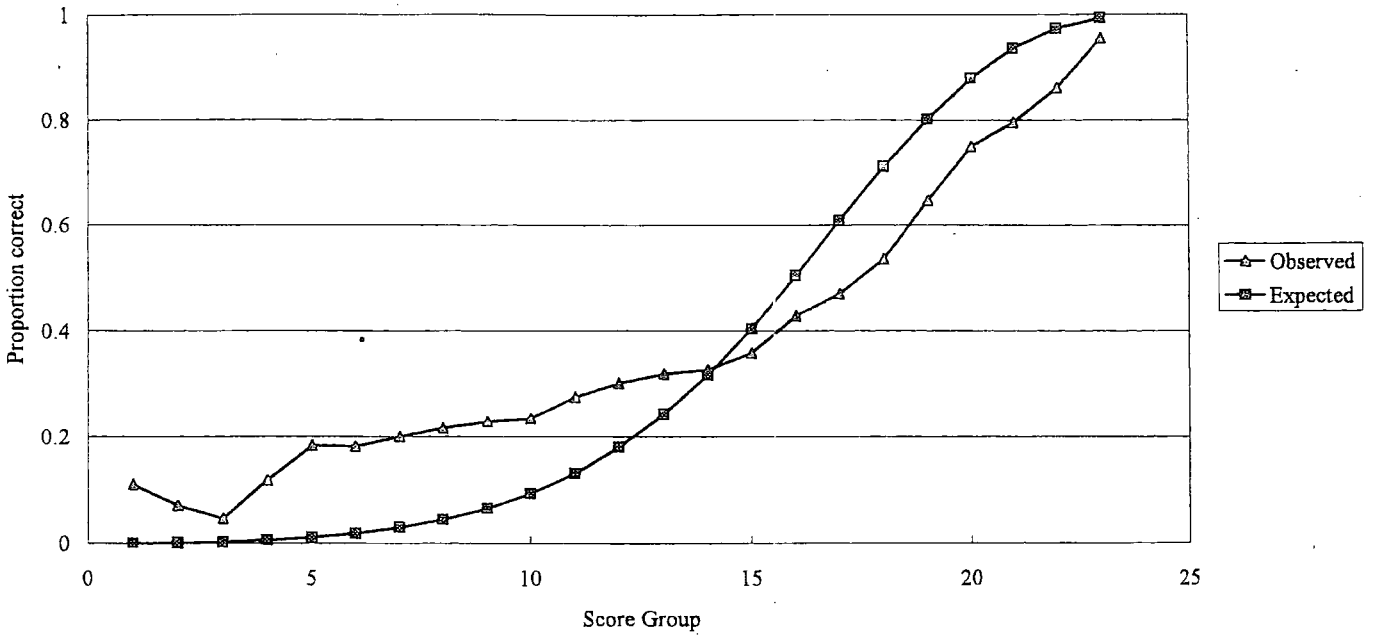
Figure 4.25
ICCs for Item 7 of Test-24
(a) Whole Group (b = 0.976)



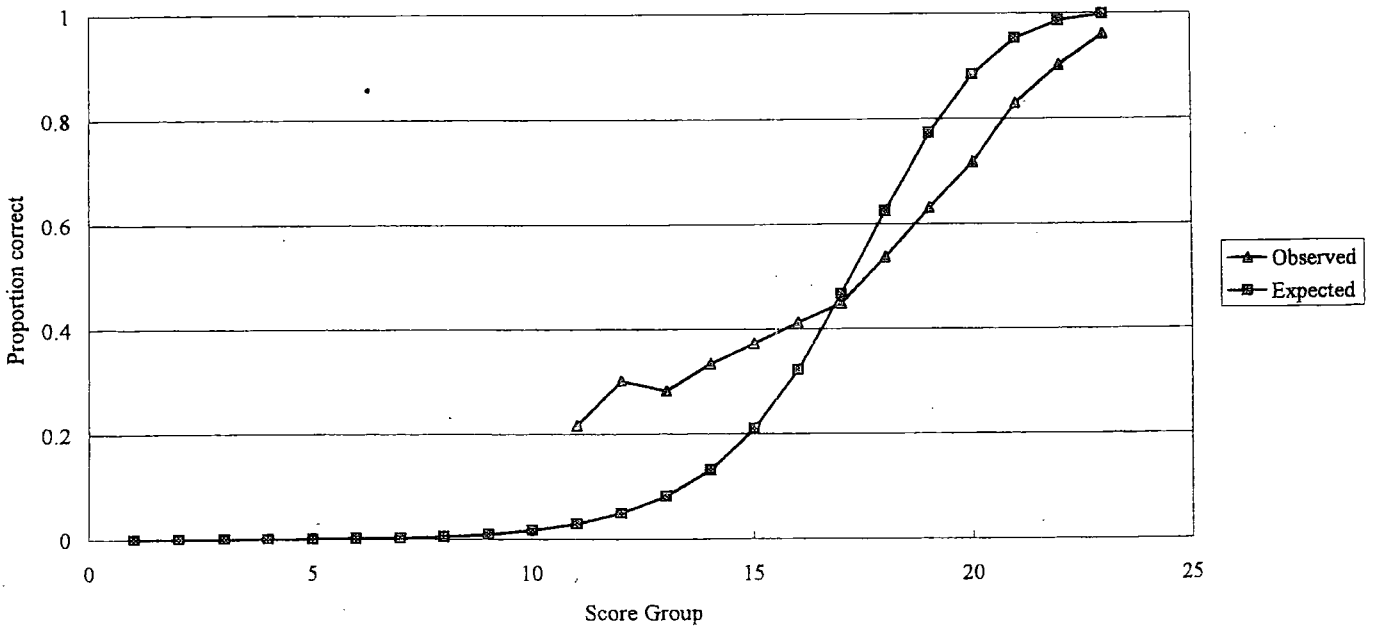
(b) Male Group (b = 0.820)



(c) Female Group (b = 1.139)



(d) High-ability Group (b = 0.078)



(e) Low-ability Group (b = 2.691)

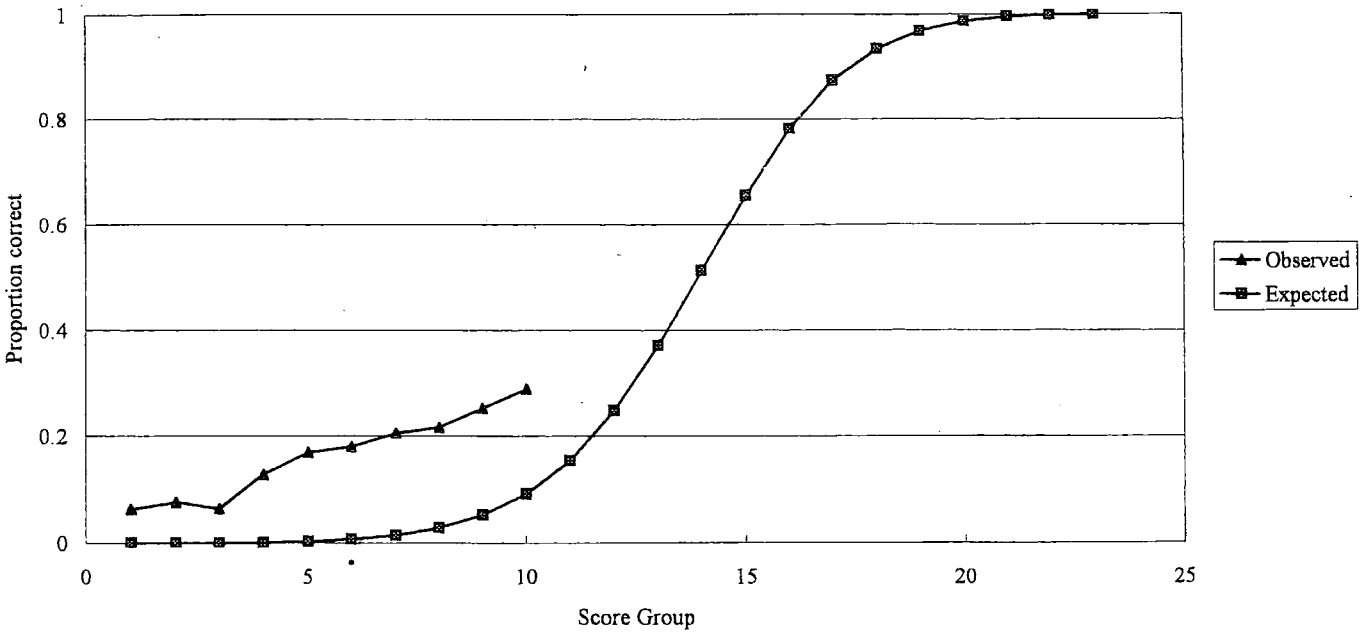
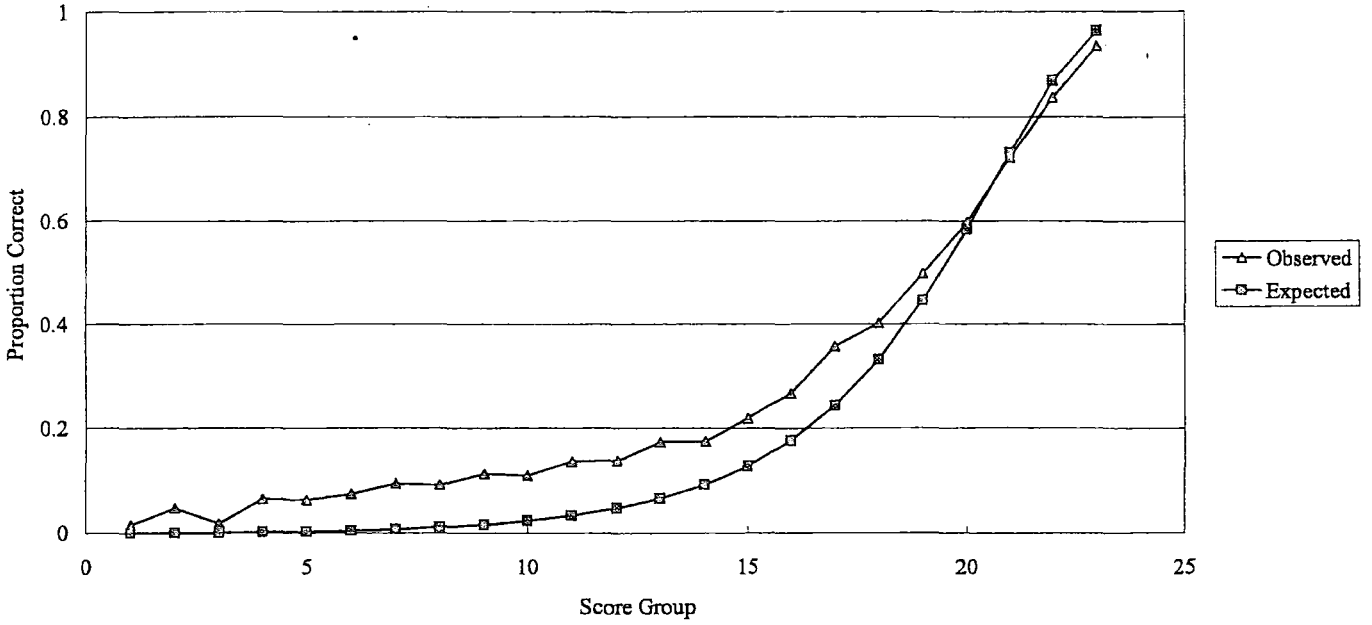
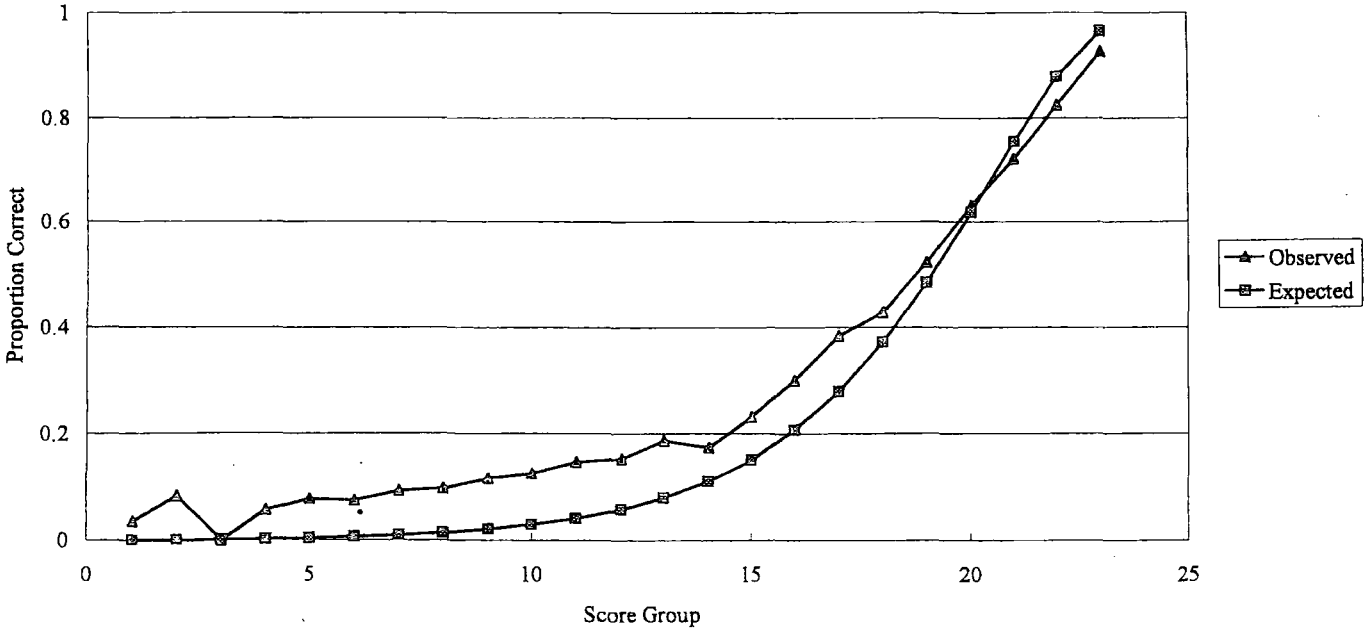


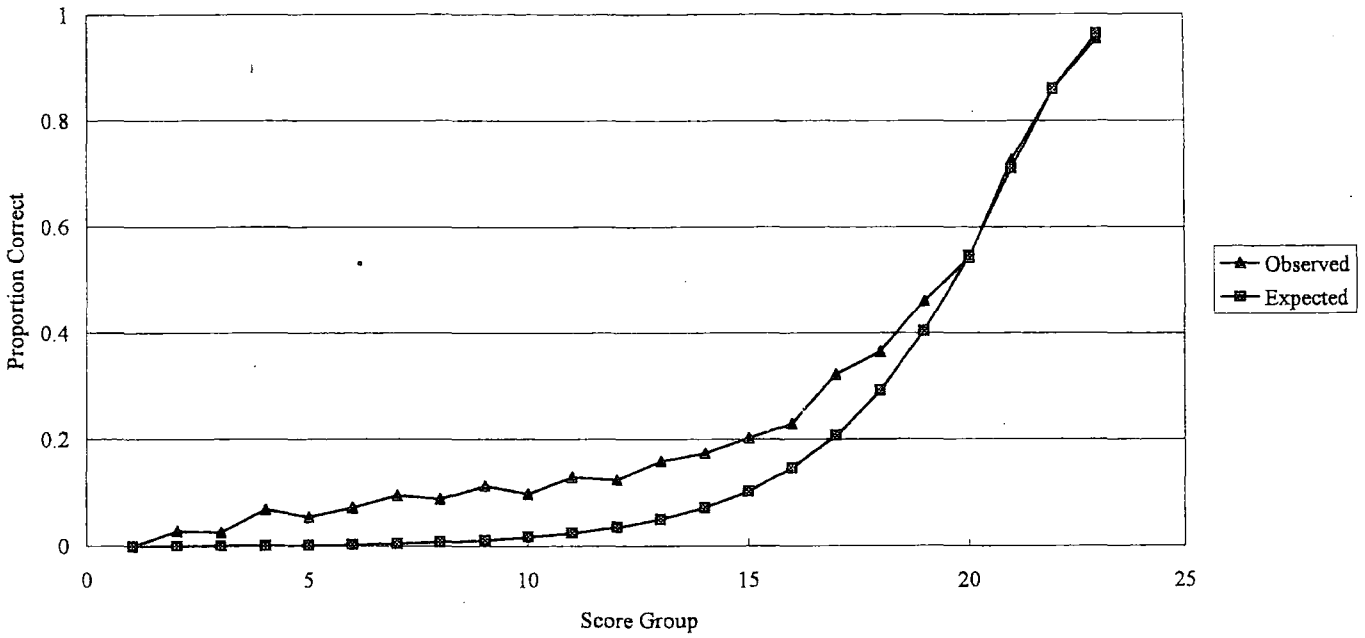
Figure 4.26
ICCs for Item 13 of Test-24
(a) Whole Group (b = 1.879)



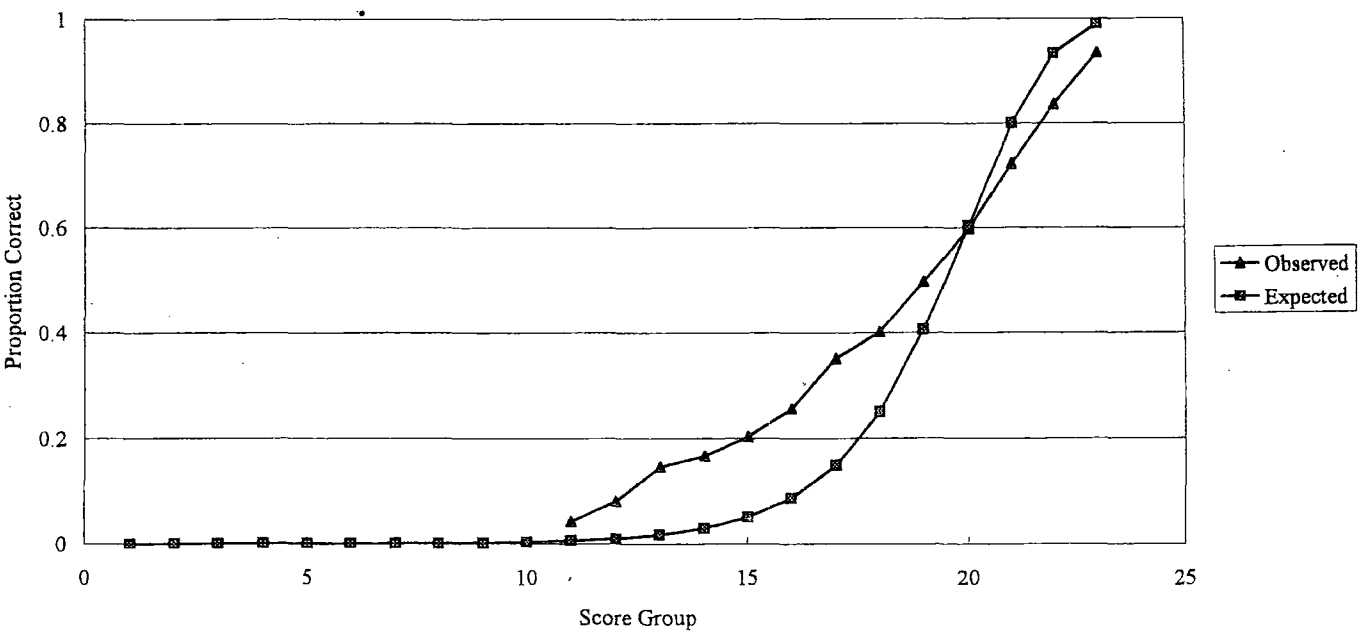
(b) Male Group (b = 1.598)



(c) Female Group (b = 2.190)



(d) High-ability Group (b = 1.023)



(e) Low-ability Group (b = 4.409)

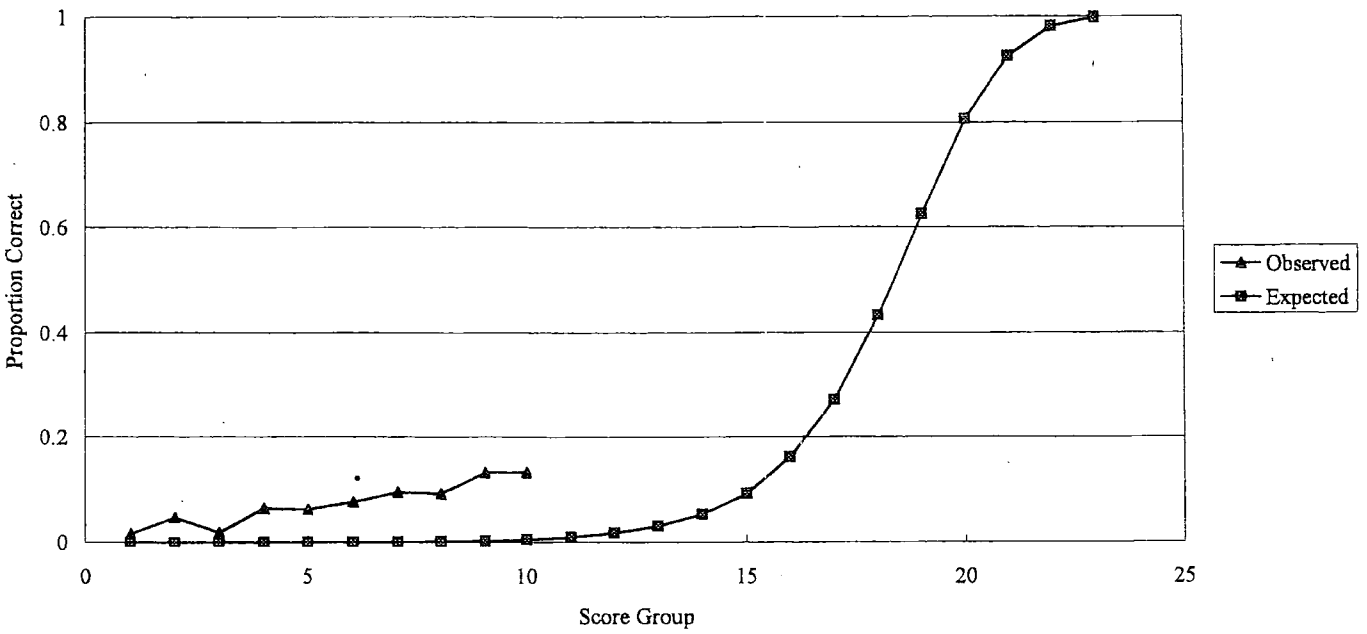
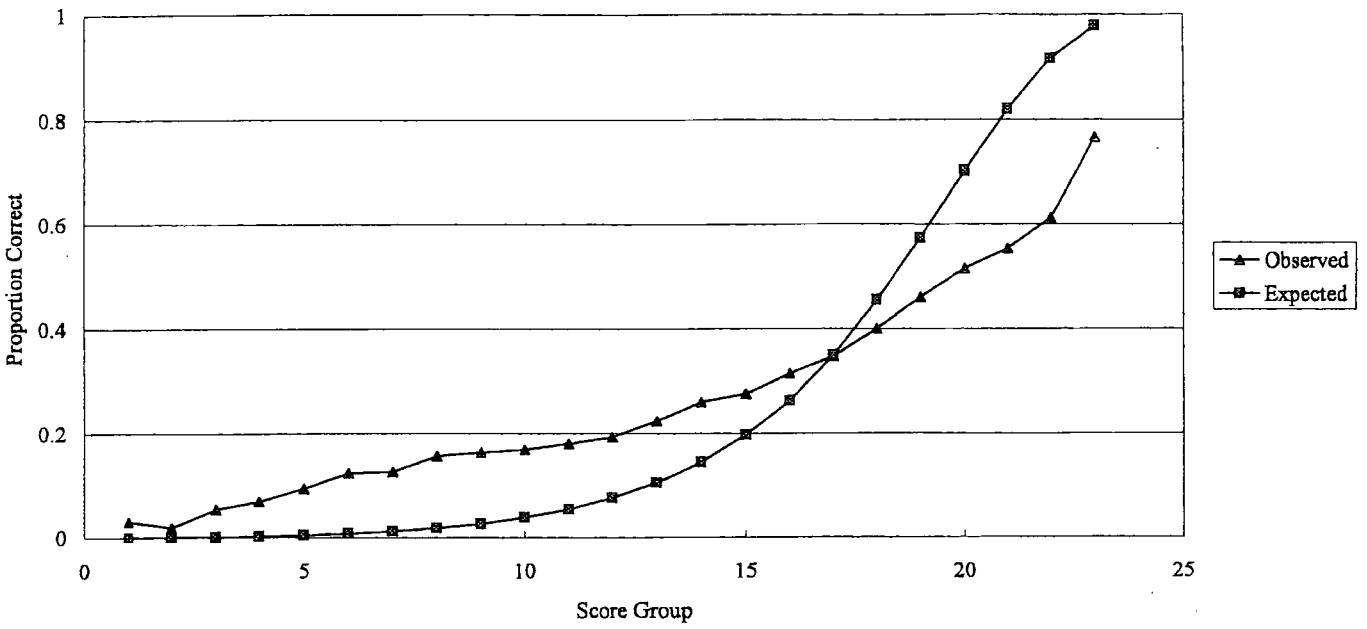
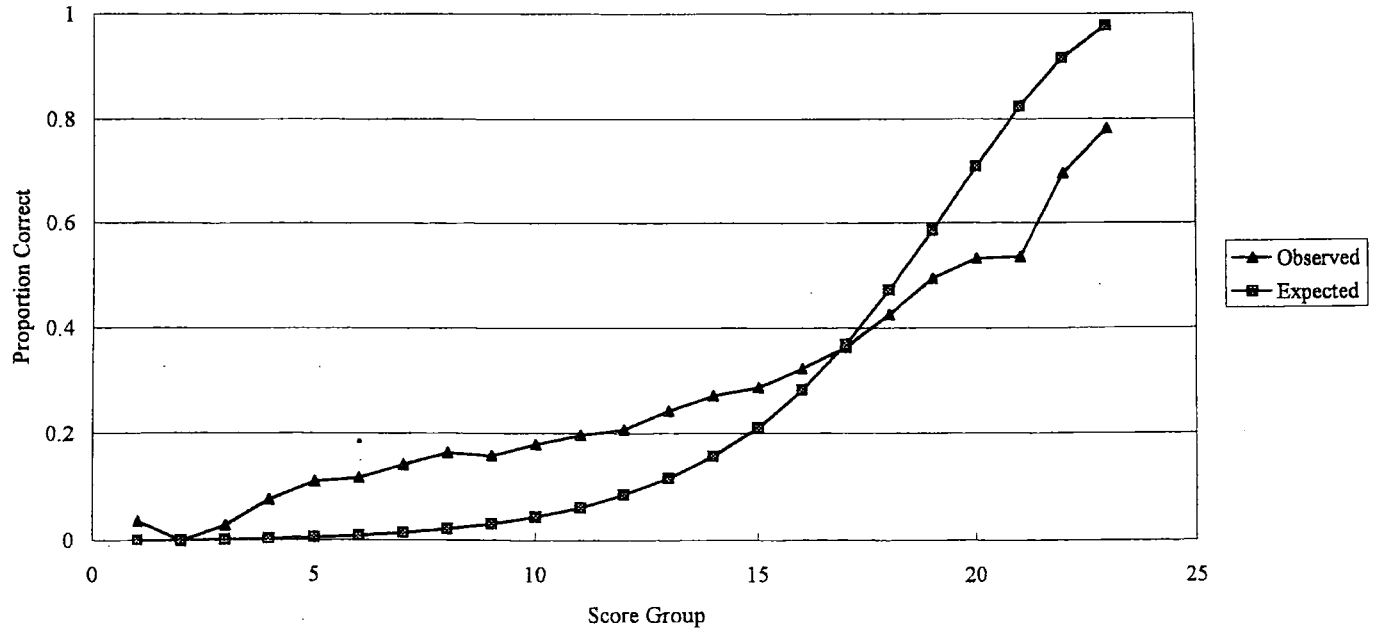


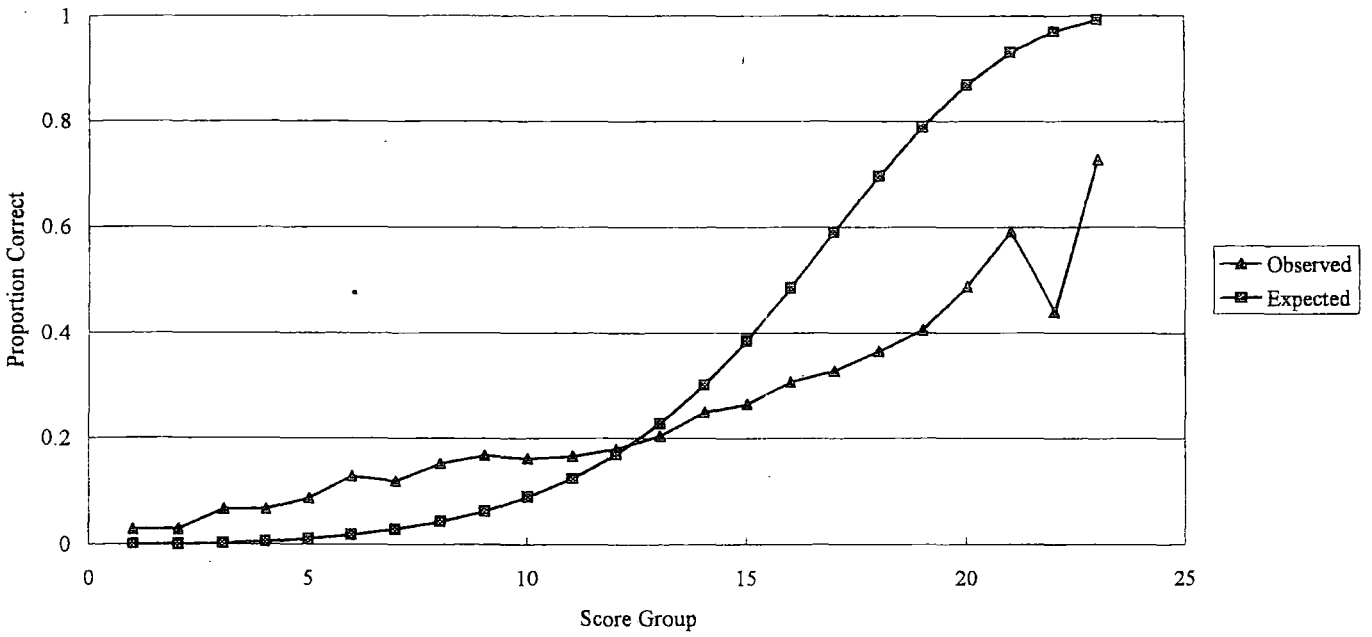
Figure 4.27
ICCs for Item 14 of Test-24
(a) Whole Group (b = 1.576)



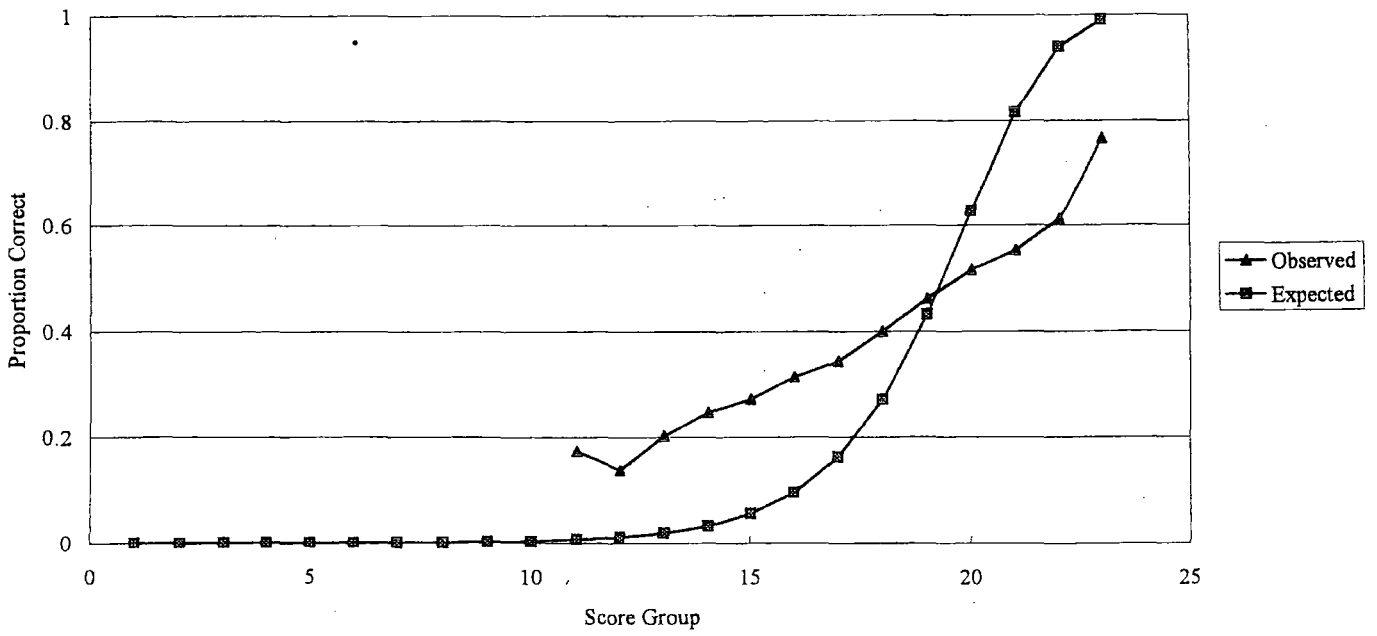
(b) Male Group (b = 1.355)



(c) Female Group (b = 1.815)



(d) High-ability Group (b = 0.962)



(e) Low-ability Group ($b = 3.669$)

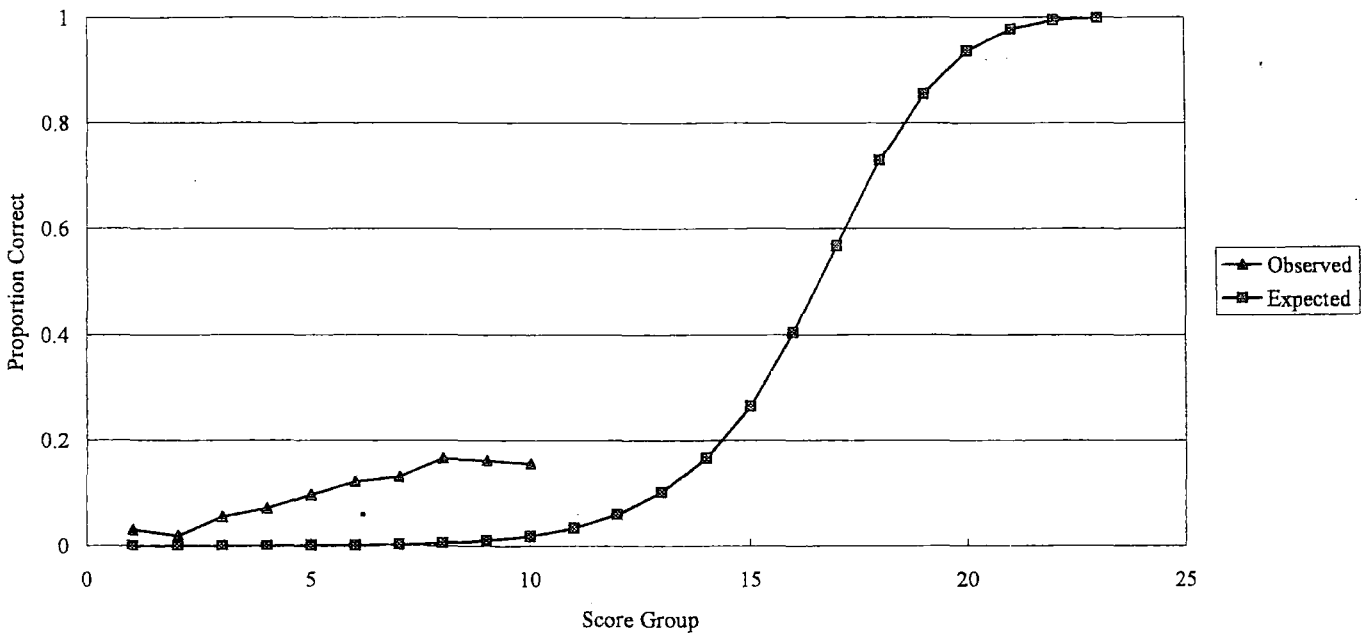
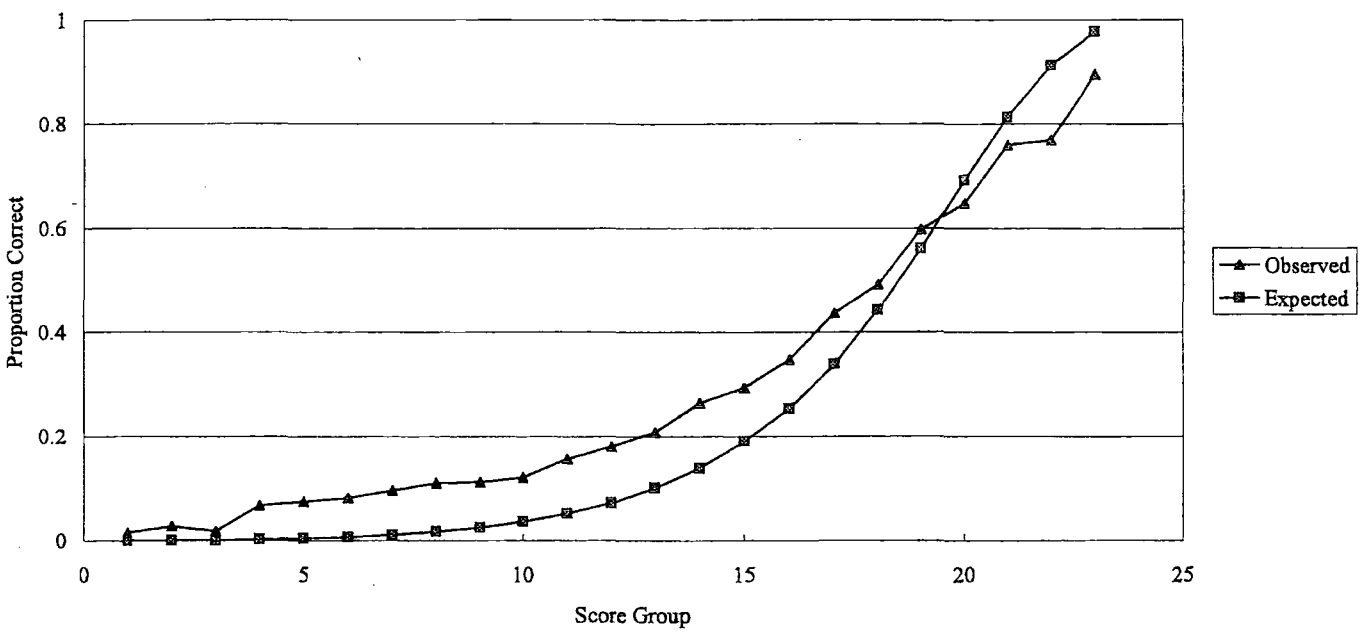
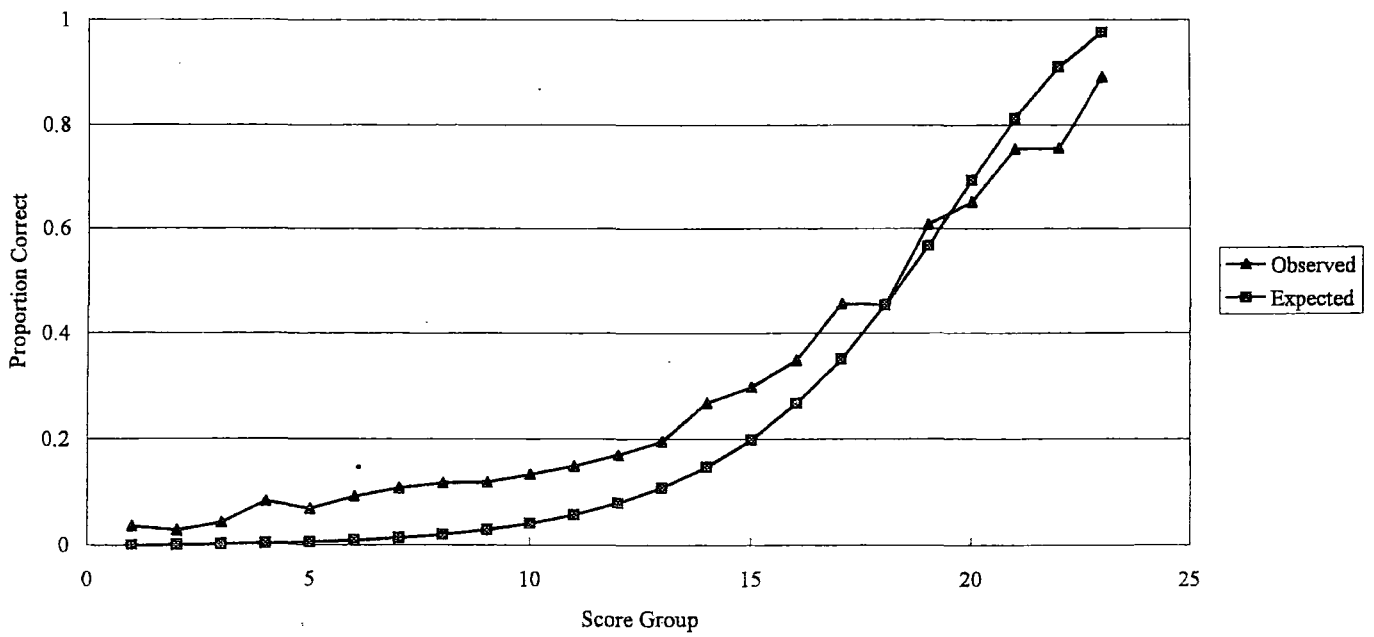


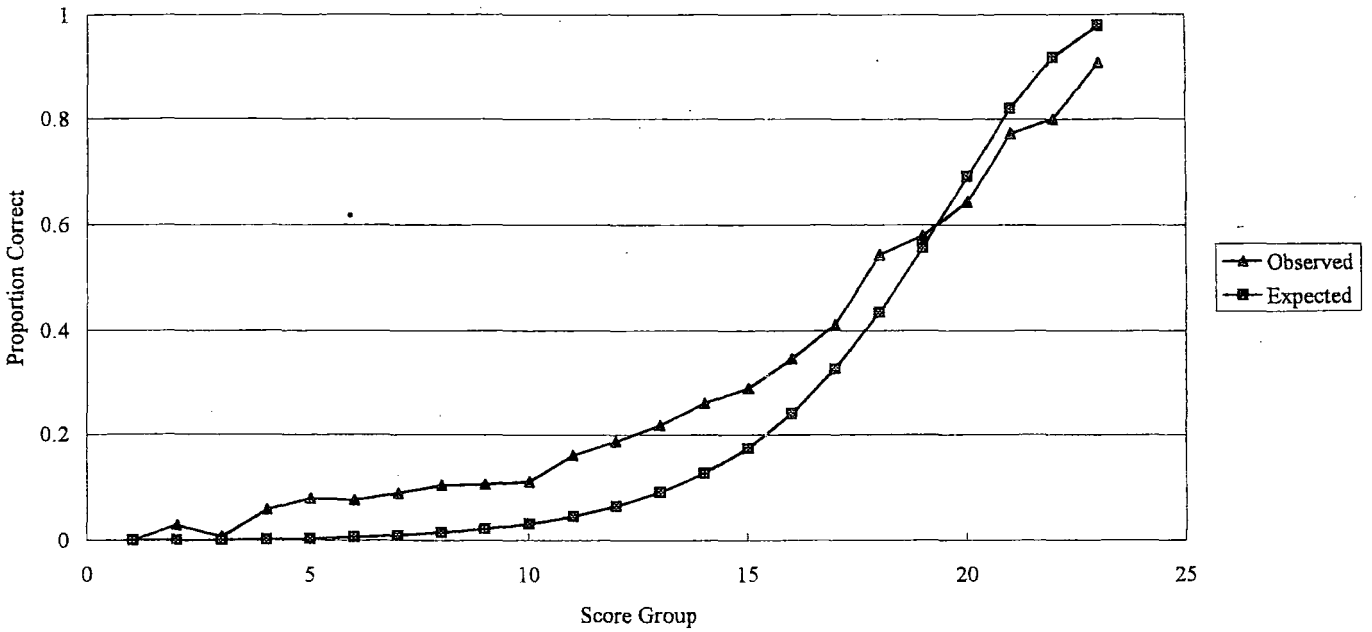
Figure 4.28
ICCs for Item 15 of Test-24
(a) Whole Group (b = 1.604)



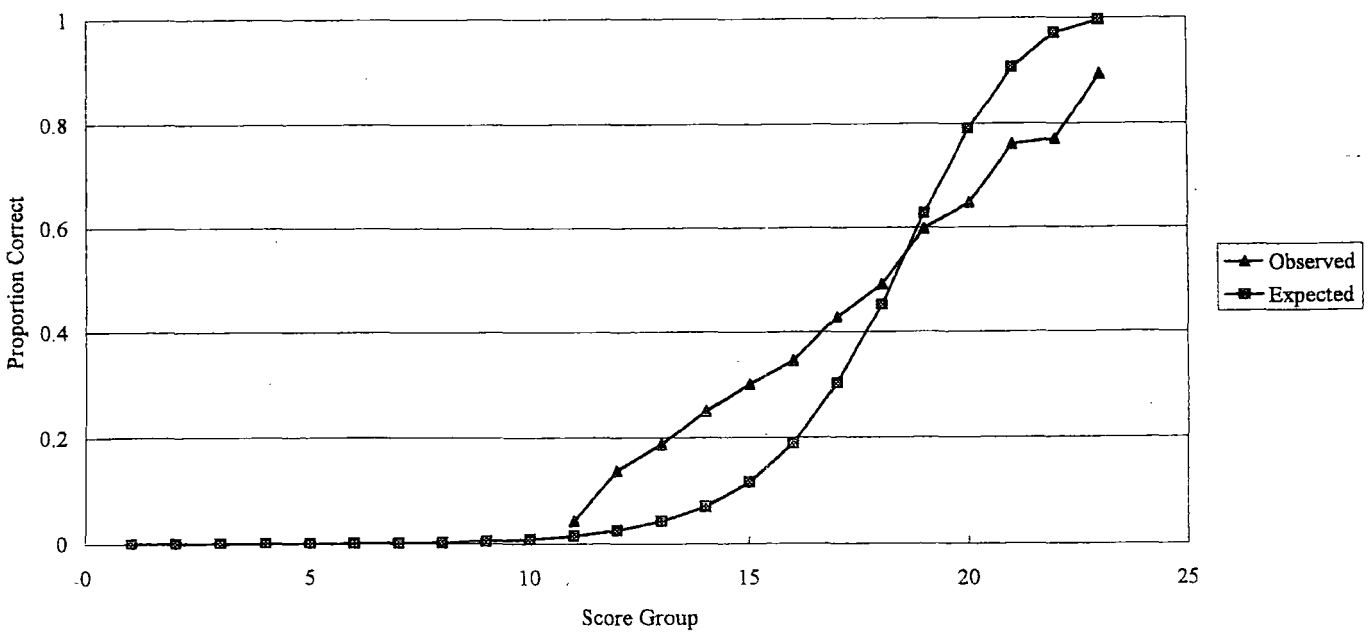
(b) Male Group (b = 1.401)



(c) Female Group (b = 1.824)



(d) High-ability Group (b = 0.492)



(e) Low-ability Group (b = 4.305)

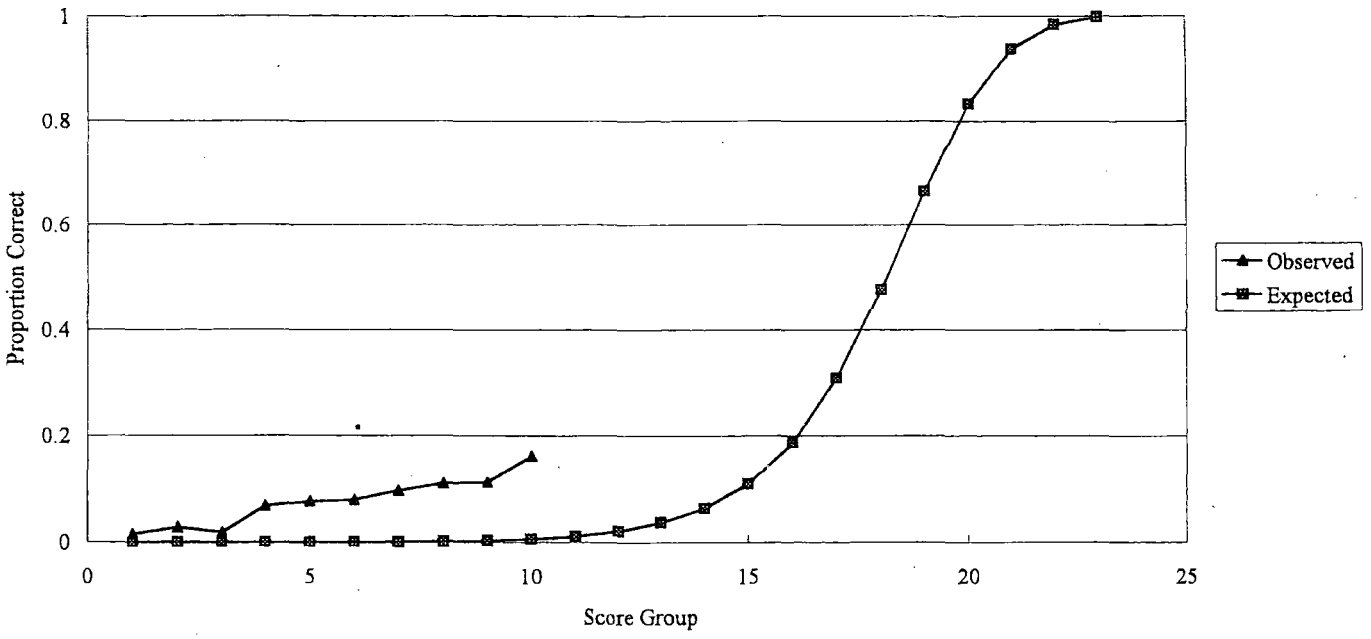
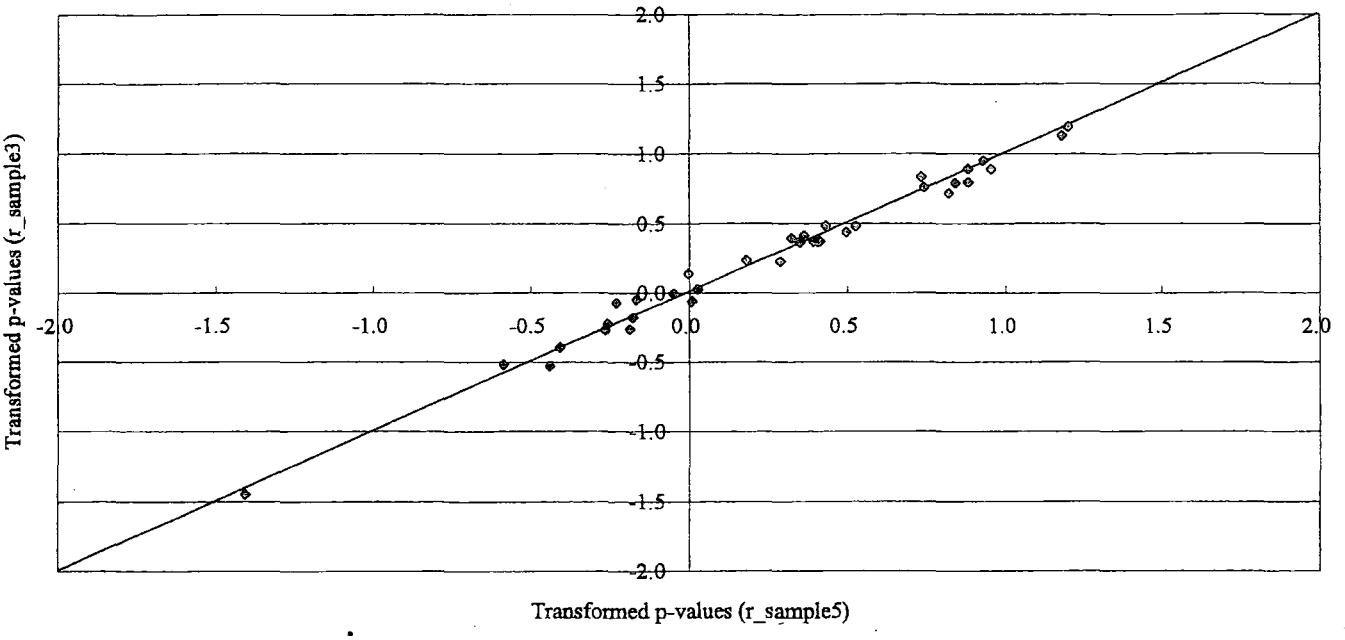


Figure 4.29
Scatterplot of CTT Transformed p-values (r_sample3 vs. r_sample5)
(a) Test-35 (r = 0.993)



(b) Test-24 (r = 0.990)

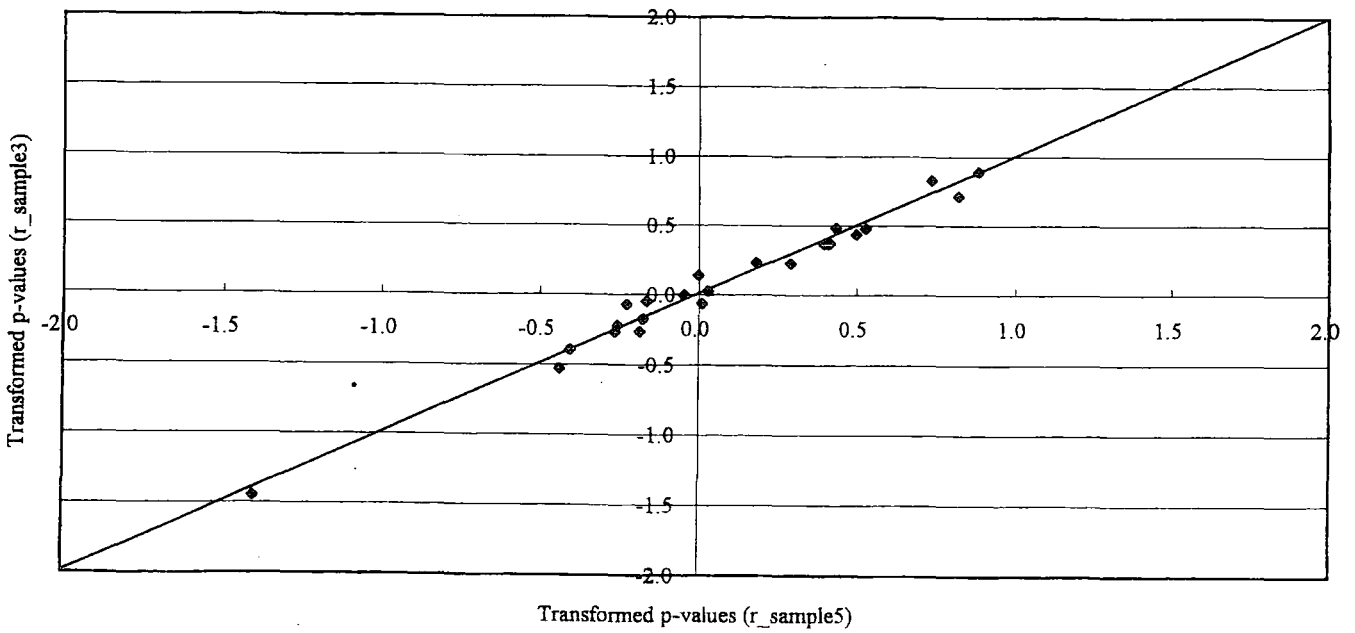
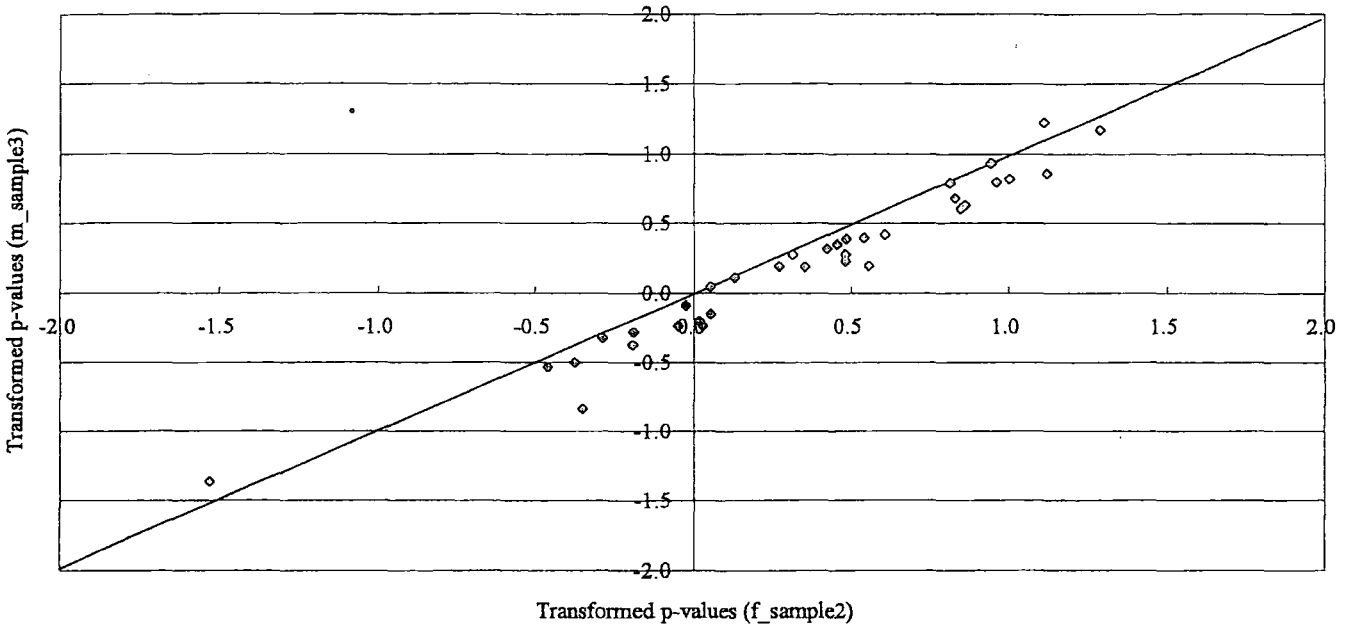


Figure 4.30
Scatterplot of CTT Transformed p-values (m_sample3 vs. f_sample2)
(a) Test-35 (r = 0.977)



(b) Test-24 (r = 0.990)

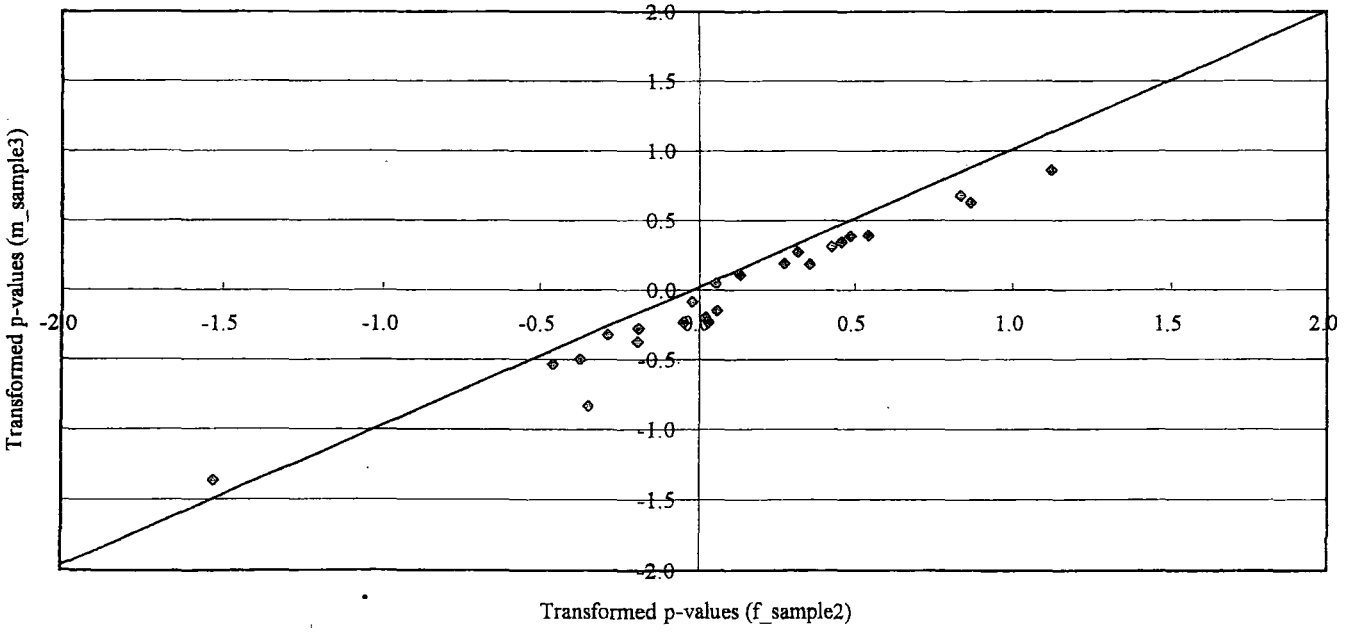
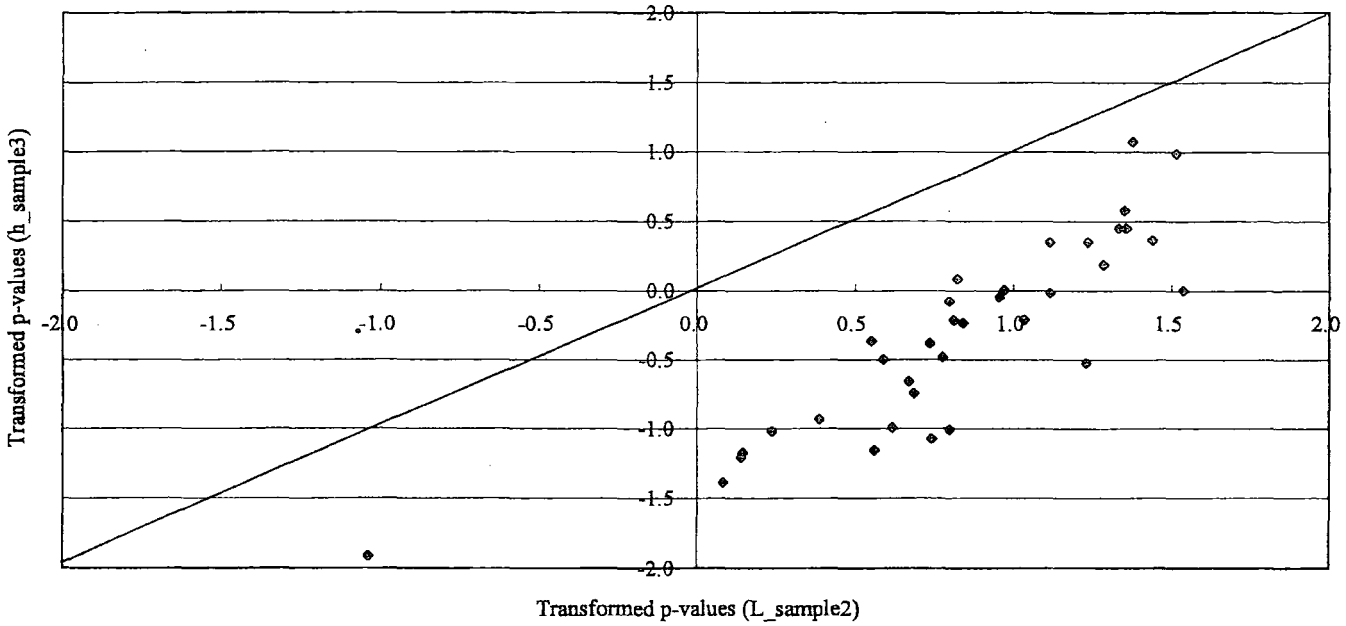


Figure 4.31
Scatterplot of CTT Transformed p-values (h_sample3 vs. L_sample2)
(a) Test-35 (r = 0.869)



(b) Test-24 (r = 0.855)

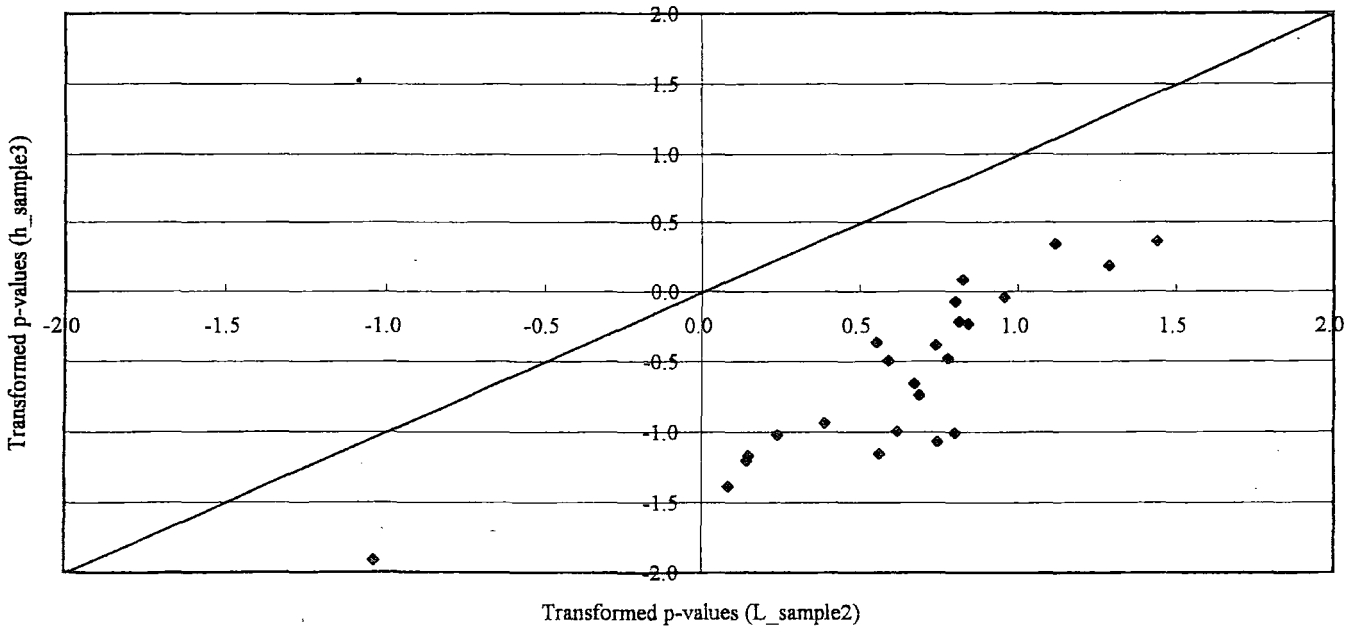
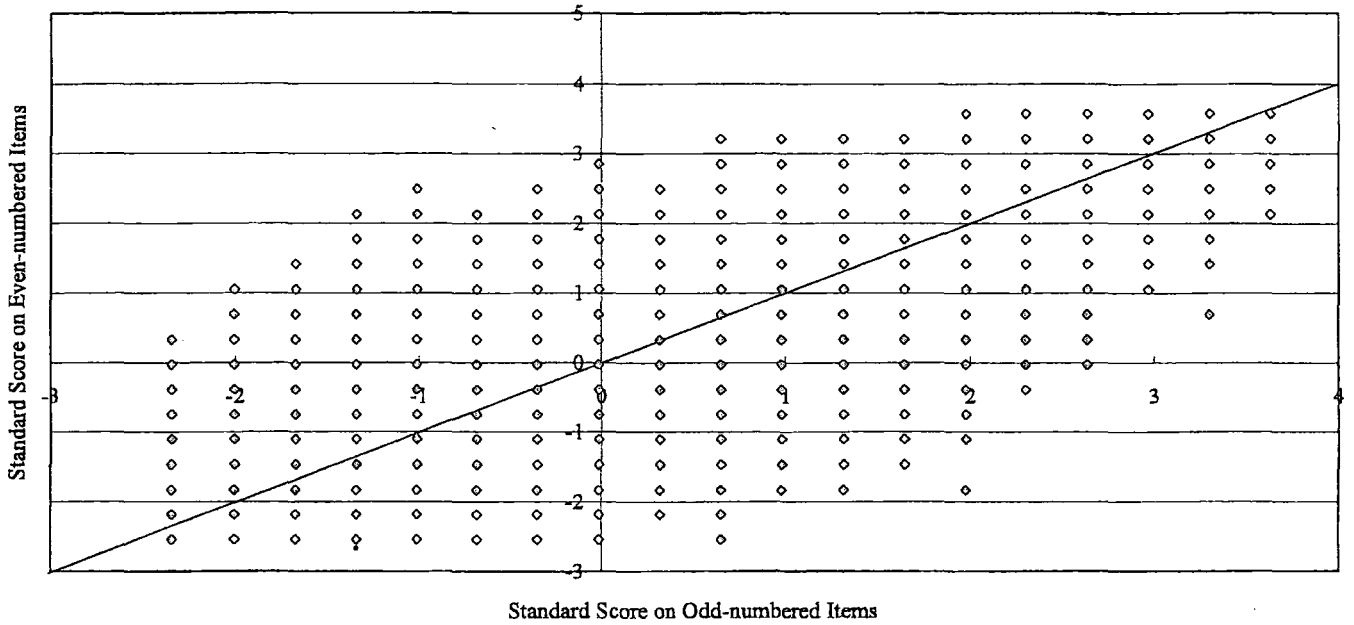


Figure 4.32
Scatterplot of CTT Ability Estimates based on Equivalent Halves
(a) Test-35 (r = 0.644)



(b) Test-24 (r = 0.589)

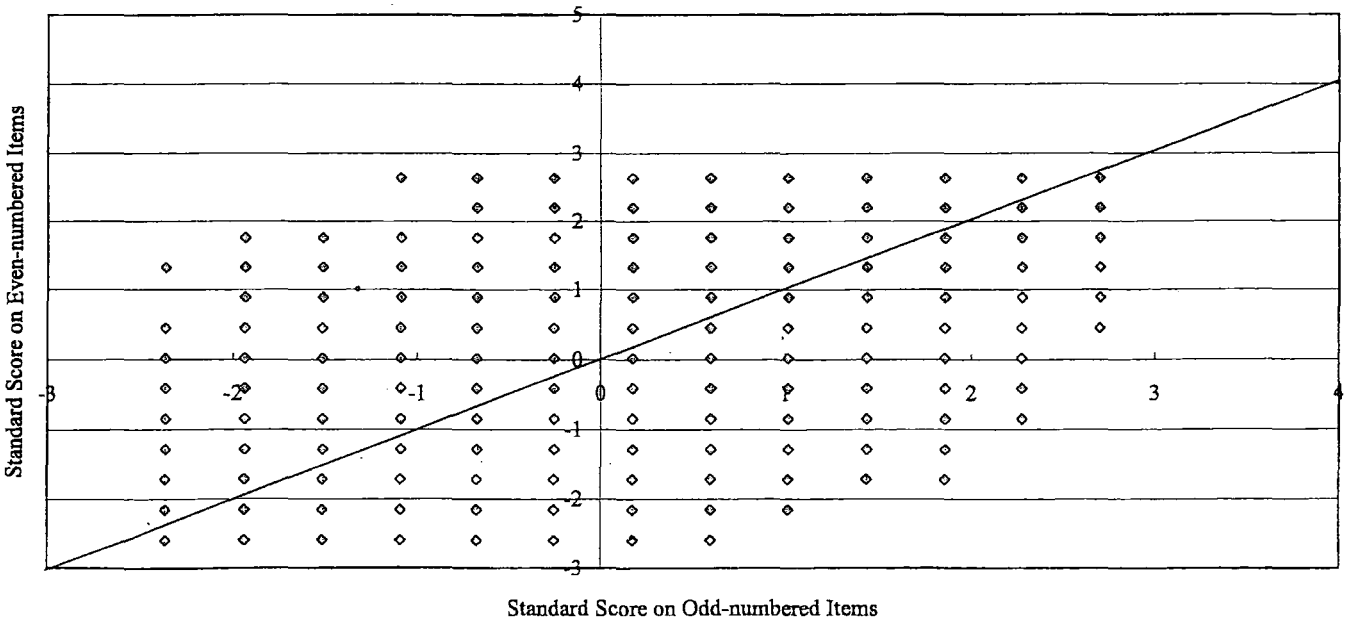
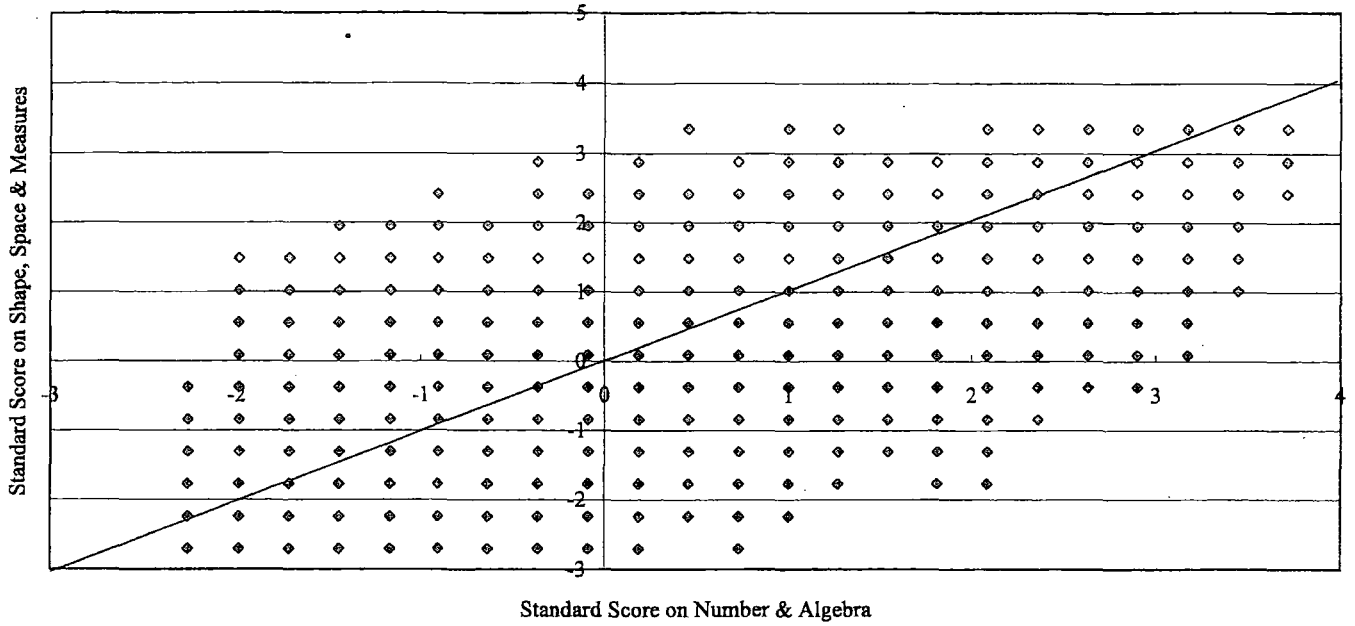


Figure 4.33
Scatterplot of CTT Ability Estimates based on Different Content Categories
(a) Test-35 (r = 0.595)



(b) Test-24 (r = 0.560)

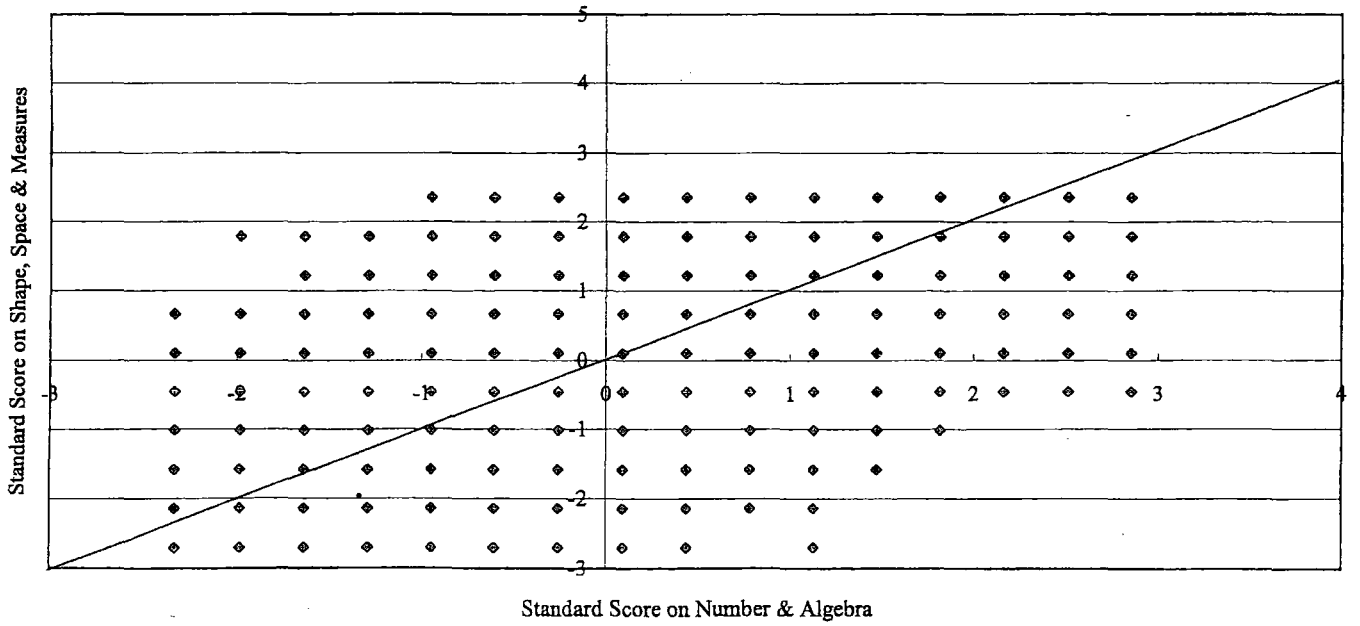
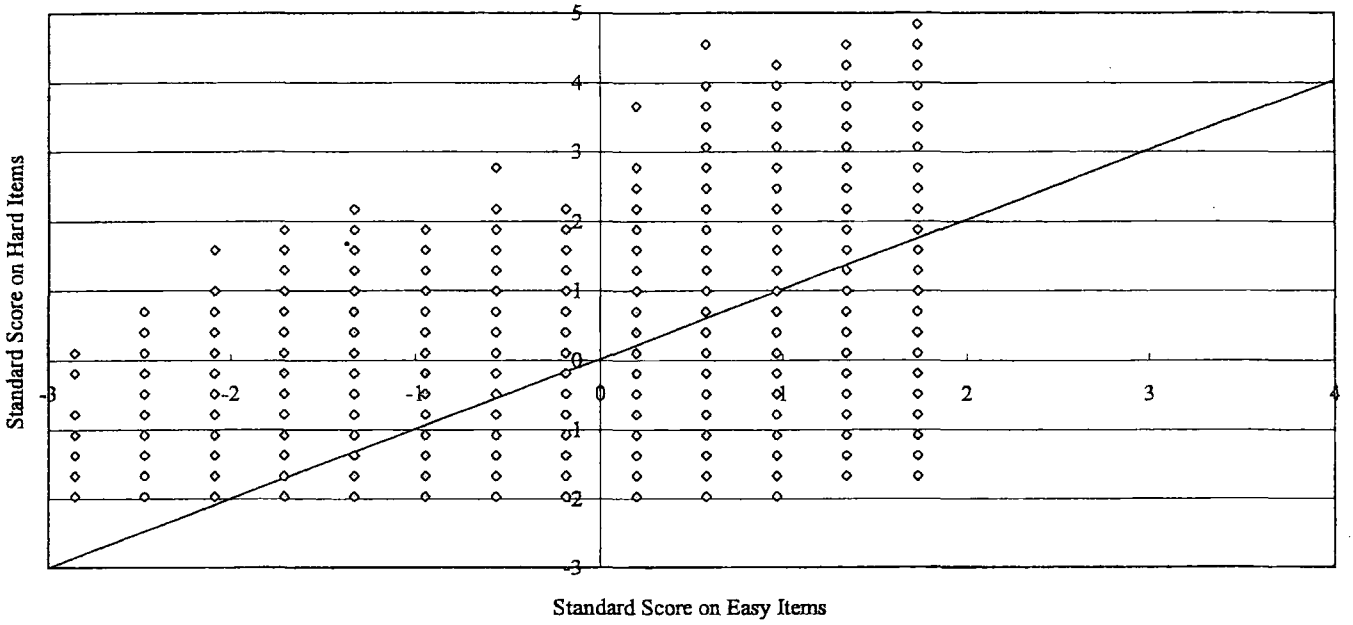


Figure 4.34
Scatterplot of CTT Ability Estimates based on Items of Different Difficulties
(a) Test-35 (r = 0.529)



(b) Test-24 (r = 0.482)

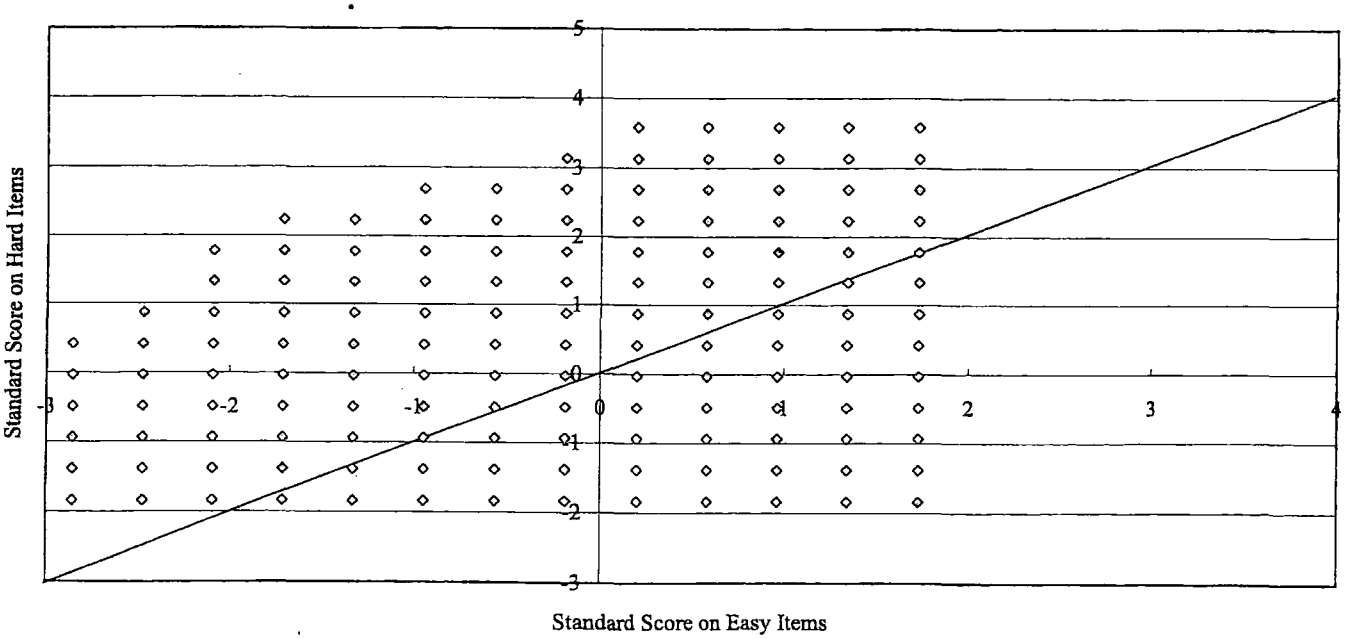
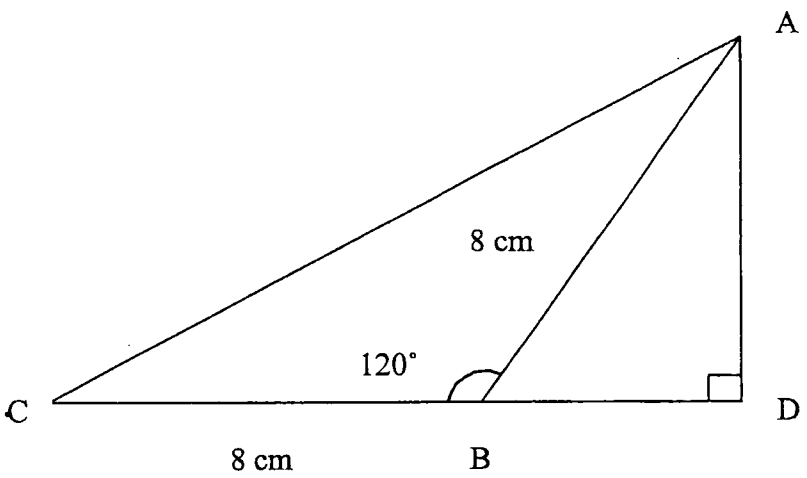


Figure 5.1

The Figure in Item 8 of the ITDA Mathematical Test



Appendix 4

List of Tables

Table 3:1

Data Sets Available for Analysis

Data Set	Size
Subject Groups:	
Whole Group	26,964
Male Group	12,567
Female Group	14,397
High-ability Group	6,601
Low-ability Group	6,821
Samples:	
Random sampling (5 samples)	1,000 each
Gender group sampling	
- Male (5 samples)	1,000 each
- Female (5 samples)	1,000 each
Ability group sampling	
- High-ability (5 samples)	1,000 each
- Low-ability (5 samples)	1,000 each

Table 3:2

Separation of the Test Items into Two Content Areas

Content Area	N&A	SSM
Item No.	1, 4, 6, 7, 9, 11, 13, 15, 16, 17, 18, 20, 21, 22, 24, 25, 26, 27, 29, 30, 32, 34	2, 3, 5, 8, 10, 12, 14, 19, 23, 28, 31, 33, 35
Total No. of Items	22	13

Table 4:1

Percentages of Examinees not Responding to Items of the Test

Item	Number of Examinees Not Responding	Percentage of Examinees Not Responding	Item	Number of Examinees Not Responding	Percentage of Examinees Not Responding
1	0	0.00	19	2,127	7.89
2	37	0.14	20	2,107	7.81
3	502	1.86	21	2,747	10.19
4	1,337	4.96	22	3,208	11.90
5	790	2.93	23	3,378	12.53
6	773	2.87	24	4,192	15.55
7	3,461	12.84	25	5,423	20.11
8	1,702	6.31	26	5,931	22.00
9	691	2.56	27	5,736	21.27
10	1,871	6.94	28	5,811	21.55
11	2,396	8.89	29	7,457	27.66
12	465	1.72	30	6,884	25.53
13	1,125	4.17	31	8,190	30.37
14	2,421	8.98	32	8,898	33.00
15	2,710	10.05	33	9,308	34.52
16	1,170	4.34	34	10,139	37.60
17	1,782	6.61	35	10,073	37.36
18	1,992	7.39	Total	26,964	

Table 4:2

Correlation Coefficients among the Item Difficulty Estimates of the 5 Random Samples in Test-35 and Test-24

Correlation Coefficients	r_sample1	r_sample2	r_sample3	r_sample4	r_sample5
r_sample1	1.000	0.995** (0.996**)	0.994** (0.995**)	0.995** (0.995**)	0.994** (0.993**)
r_sample2	0.995** (0.996**)	1.000	0.997** (0.995**)	0.996** (0.995**)	0.995** (0.993**)
r_sample3	0.994** (0.995**)	0.997** (0.995**)	1.000	0.995** (0.993**)	0.993** (0.990**)
r_sample4	0.995** (0.995**)	0.996** (0.995**)	0.995** (0.993**)	1.000	0.999** (0.998**)
r_sample5	0.994** (0.993**)	0.995** (0.993**)	0.993** (0.990**)	0.999** (0.998**)	1.000

** Correlation is statistically significant at the 0.01 level (2-tailed).

Note: The correlation coefficients for Test-24 are presented within parentheses.

Table 4:3

Correlation Coefficients between the Item Difficulty Estimates of the Male and Female Samples in Test-35 and Test-24

Correlation Coefficients	f_sample1	f_sample2	f_sample3	f_sample4	f_sample5
m_sample1	0.982 (0.973)	0.979 (0.971)	0.982 (0.971)	0.979 (0.973)	0.983 (0.976)
m_sample2	0.985 (0.980)	0.982 (0.978)	0.984 (0.975)	0.981 (0.979)	0.986 (0.983)
m_sample3	0.982 (0.976)	0.976 (0.974)	0.982 (0.972)	0.976 (0.974)	0.983 (0.981)
m_sample4	0.983 (0.976)	0.979 (0.973)	0.983 (0.974)	0.980 (0.978)	0.984 (0.978)
m_sample5	0.986 (0.977)	0.983 (0.976)	0.984 (0.975)	0.983 (0.975)	0.988 (0.982)

Notes: (1) All correlations are statistically significant at the 0.01 level (2-tailed).

(2) The correlation coefficients for Test-24 are presented within parentheses.

Table 4:4

Correlation Coefficients between the Item Difficulty Estimates of the High-ability and Low-ability Samples in Test-35 and Test-24

Correlation Coefficients	L_sample1	L_sample2	L_sample3	L_sample4	L_sample5
h_sample1	0.900 (0.892)	0.894 (0.889)	0.897 (0.894)	0.892 (0.896)	0.910 (0.894)
h_sample2	0.895 (0.879)	0.885 (0.873)	0.889 (0.876)	0.886 (0.881)	0.904 (0.878)
h_sample3	0.888 (0.875)	0.882 (0.872)	0.885 0.876)	0.883 (0.881)	0.902 (0.879)
h_sample4	0.895 (0.874)	0.889 (0.870)	0.892 (0.875)	0.888 (0.876)	0.906 (0.876)
h_sample5	0.896 (0.894)	0.892 (0.890)	0.893 (0.891)	0.890 (0.897)	0.908 (0.894)

Notes: (1) All correlations are statistically significant at the 0.01 level (2-tailed).

(2) The correlation coefficients for Test-24 are presented within parentheses.

Table 4:5

Comparability of the Invariance of Item Difficulty Estimates in Different Sampling Plans

Sampling Plan	Number of Correlation Coefficients	Average Correlation Coefficients*	
		Test-35	Test-24
Random sampling	10	0.996	0.995
Gender group sampling	25	0.982	0.976
Ability group sampling	25	0.894	0.883

* The average correlation coefficients are computed as described in para. 3.4.

Table 4:6

Unfitted Items in Various Samples

Sampling Plan	Sample	Test-35		Test-24	
		Unfitted items in each sample	"Really" unfitted items*	Unfitted items in each sample	"Really" unfitted items*
Random sampling	r_sample1	7, 10, 20, 21, 26, 31, 32, 33, 34, 35	7, 12, 19, 20, 21, 27,	3, 7, 14, 20, 23, 24	3, 7, 11, 14, 20, 24
	r_sample2	3, 12, 13, 19, 20, 21, 26, 31, 32, 33, 34, 35	31, 33, 34, 35	3, 10, 11, 13, 24	
	r_sample3	7, 12, 20, 21, 27, 31, 34		6, 7, 11, 14	
	r_sample4	4, 7, 11, 12, 19, 21, 23, 27, 31, 33, 34, 35		7, 11, 14, 19, 20, 24	
	r_sample5	3, 7, 14, 19, 20, 21, 23, 27, 31, 33, 34, 35	(28.6%)	3, 7, 10, 11, 14, 19, 20, 24	(25.0%)
Gender group sampling	m_sample1	3, 9, 20, 26, 27, 31, 33, 34, 35	7, 11, 19, 20, 21, 27,	7, 10, 14, 20, 24	7, 10, 11, 20, 24
	m_sample2	5, 6, 7, 11, 19, 21, 27, 31, 33, 34, 35	31, 33, 34, 35	3, 5, 7, 10, 11	
	m_sample3	3, 7, 11, 14, 18, 19, 20, 21, 27, 32, 34		3, 7, 11, 20	
	m_sample4	1, 11, 18, 19, 20, 21, 22, 24, 26, 27, 31, 32, 33, 34, 35		10, 11, 12, 19, 20, 21, 24	
	m_sample5	7, 10, 20, 24, 27, 31, 33, 34, 35	(28.6%)	7, 14, 20, 22, 24	(20.8%)
	f_sample1	6, 13, 14, 19, 28, 30, 32, 34	7, 14, 19, 27, 34	14, 20, 24	14, 20, 24
	f_sample2	7, 14, 19, 21, 24, 34		7, 20, 24	
	f_sample3	7, 19, 20, 21, 27, 31, 32, 34, 35		3, 13, 14, 21, 23, 24	
	f_sample4	6, 7, 14, 19, 20, 22, 27, 31		7, 11, 14, 20, 24	
	f_sample5	1, 4, 13, 14, 17, 24, 27, 34	(14.3%)	13, 14, 22, 24	(12.5%)
Ability group sampling	h_sample1	10, 22, 23, 25	22, 23	23	22, 23
	h_sample2	19, 22, 31, 34		22, 23	
	h_sample3	2, 3, 4, 22, 23		2, 22, 23, 24	
	h_sample4	3, 21, 23, 31, 35		5, 14, 23	
	h_sample5	9, 10, 19, 20, 23, 27, 34	(5.7%)	22	(8.3%)
	L_sample1	1, 3, 6, 13	1, 3	/	/
	L_sample2	23		/	
	L_sample3	1, 3, 4, 13		/	
	L_sample4	1, 3, 7, 11		6, 9, 24	
	L_sample5	1, 4	(5.7%)	8, 9, 13	(0%)

* The "really" unfitted items are items which are found unfitted in at least 3 samples in each sampling group.

Note: The percentages of "really" unfitted items are presented within parentheses.

Table 4:7

Effect Sizes of Individual Items of Various Subject Groups

Item difficulty estimates						Effect Sizes			
Item	Whole Group	Male Group	Female Group	High-ability Group	Low-ability Group	Male Group	Female Group	High-ability Group	Low-ability Group
1	-1.213	-1.752	-0.811	-4.686	0.055	-0.430	0.321	-2.774	1.013
2	-3.287	-3.169	-3.468	-7.139	-3.538	0.094	-0.145	-3.076	-0.200
3	0.257	0.148	0.365	-1.163	1.991	-0.087	0.086	-1.134	1.385
4	0.563	0.413	0.715	-1.340	2.753	-0.120	0.121	-1.520	1.749
5	-0.910	-1.029	-0.820	-3.853	0.452	-0.095	0.072	-2.350	1.088
6	0.591	0.498	0.691	-1.329	2.881	-0.074	0.080	-1.533	1.829
7	1.051	0.885	1.229	0.076	2.924	-0.133	0.142	-0.779	1.496
8	-1.036	-1.171	-0.935	-4.034	0.428	-0.108	0.081	-2.394	1.169
9	-0.383	-0.547	-0.239	-2.864	1.320	-0.131	0.115	-1.981	1.360
10	0.700	0.598	0.810	-0.588	2.669	-0.081	0.088	-1.029	1.573
11	0.865	0.795	0.949	-0.326	2.749	-0.056	0.067	-0.951	1.505
12	0.788	0.644	0.940	-0.627	2.855	-0.115	0.121	-1.130	1.651
13	2.015	1.720	2.355	1.173	4.796	-0.236	0.272	-0.672	2.221
14	1.691	1.459	1.953	1.102	3.989	-0.185	0.209	-0.470	1.835
15	1.721	1.509	1.962	0.556	4.683	-0.169	0.192	-0.930	2.366
16	-0.648	-0.834	-0.492	-3.141	1.040	-0.149	0.125	-1.991	1.348
17	0.109	-0.016	0.228	-1.668	2.085	-0.100	0.095	-1.419	1.578
18	-0.437	-0.565	-0.327	-3.208	1.858	-0.102	0.088	-2.213	1.833
19	-0.585	-0.697	-0.493	-3.723	1.877	-0.089	0.073	-2.506	1.966
20	-0.264	-0.543	-0.016	-3.567	2.607	-0.223	0.198	-2.638	2.293
21	-0.103	-0.391	0.159	-3.223	2.800	-0.230	0.209	-2.492	2.318
22	-0.096	-0.197	-0.004	-2.047	2.142	-0.081	0.073	-1.558	1.787
23	-0.183	-0.304	-0.074	-2.353	2.315	-0.097	0.087	-1.733	1.995
24	0.863	0.752	0.985	-0.109	3.322	-0.089	0.097	-0.776	1.964
25	1.211	1.057	1.382	-0.178	4.081	-0.123	0.137	-1.109	2.292
26	1.764	1.543	2.014	1.059	4.293	-0.177	0.200	-0.563	2.020
27	0.827	0.468	1.196	-1.712	4.624	-0.287	0.295	-2.028	3.032
28	0.763	0.552	0.979	-0.729	3.687	-0.169	0.173	-1.192	2.335
29	1.983	1.768	2.232	1.296	5.114	-0.172	0.199	-0.549	2.501
30	1.631	1.376	1.915	0.113	5.442	-0.204	0.227	-1.212	3.044
31	0.857	0.634	1.087	0.061	3.607	-0.178	0.184	-0.636	2.196
32	1.856	1.652	2.091	1.422	4.732	-0.163	0.188	-0.347	2.297
33	2.060	1.918	2.235	1.894	4.919	-0.113	0.140	-0.133	2.283
34	2.640	2.492	2.833	3.336	5.303	-0.118	0.154	0.556	2.127
35	2.769	2.506	3.091	3.034	5.500	-0.210	0.257	0.212	2.181
Mean	0.584	0.405	0.763	-1.100	2.924				
SD	1.252	1.226	1.317	2.304	1.872				

Table 4:8

Item Ordering of Various Subject Groups

Item Number									
Test-35					Test-24				
Whole Group	Male Group	Female Group	High-ability Group	Low-ability Group	Whole Group	Male Group	Female Group	High-ability Group	Low-ability Group
2	2	2	2	2	2	2	2	2	2
1	1	8	1	1	1	1	8	1	1
8	8	5	8	8	8	8	5	8	8
5	5	1	5	5	5	5	1	5	5
16	16	19	19	16	16	16	19	19	16
19	19	16	20	9	19	19	16	20	9
18	18	18	21	18	18	18	18	21	18
9	9	9	18	19	9	9	9	18	19
20	20	23	16	3	20	20	23	16	3
23	21	20	9	17	23	21	20	9	17
21	23	22	23	22	21	23	22	23	22
22	22	21	22	23	22	22	21	22	23
17	17	17	27	20	17	17	17	17	20
3	3	3	17	10	3	3	3	4	10
4	4	6	4	11	4	4	6	6	11
6	27	4	6	4	6	6	4	3	4
10	6	10	3	21	10	10	10	12	21
28	28	12	28	12	12	12	12	10	12
12	10	11	12	6	24	24	11	11	6
27	31	28	10	7	11	11	24	24	7
31	12	24	11	24	7	7	7	7	24
24	24	31	25	31	14	14	14	15	14
11	11	27	24	28	15	15	15	14	15
7	7	7	31	14	13	13	13	13	13
25	25	25	7	25					
30	30	30	30	26					
14	14	14	15	27					
15	15	15	26	15					
26	26	26	14	32					
32	32	32	13	13					
29	13	29	29	33					
13	29	33	32	29					
33	33	13	33	34					
34	34	34	35	30					
35	35	35	34	35					

Table 4:9

Item by Item Difficulty Estimate Distribution

Test-35				Test-24			
Items	Item Difficulty Estimates	Items (in item difficulty order)	Item Difficulty Estimates	Items	Item Difficulty Estimates	Items (in item difficulty order)	Item Difficulty Estimates
1	-1.213	2	-3.287	1	-1.133	2	-3.035
2	-3.287	1	-1.213	2	-3.035	1	-1.133
3	0.257	8	-1.036	3	0.234	8	-0.969
4	0.563	5	-0.910	4	0.520	5	-0.852
5	-0.910	16	-0.648	5	-0.852	16	-0.608
6	0.591	19	-0.585	6	0.546	19	-0.550
7	1.051	18	-0.437	7	0.976	18	-0.412
8	-1.036	9	-0.383	8	-0.969	9	-0.362
9	-0.383	20	-0.264	9	-0.362	20	-0.252
10	0.700	23	-0.183	10	0.648	23	-0.176
11	0.865	21	-0.103	11	0.803	21	-0.102
12	0.788	22	-0.096	12	0.731	22	-0.096
13	2.015	17	0.109	13	1.879	17	0.096
14	1.691	3	0.257	14	1.576	3	0.234
15	1.721	4	0.563	15	1.604	4	0.520
16	-0.648	6	0.591	16	-0.608	6	0.546
17	0.109	10	0.700	17	0.096	10	0.648
18	-0.437	28	0.763	18	-0.412	12	0.731
19	-0.585	12	0.788	19	-0.550	24	0.801
20	-0.264	27	0.827	20	-0.252	11	0.803
21	-0.103	31	0.857	21	-0.102	7	0.976
22	-0.096	24	0.863	22	-0.096	14	1.576
23	-0.183	11	0.865	23	-0.176	15	1.604
24	0.863	7	1.051	24	0.801	13	1.879
25	1.211	25	1.211				
26	1.764	30	1.631				
27	0.827	14	1.691				
28	0.763	15	1.721				
29	1.983	26	1.764				
30	1.631	32	1.856				
31	0.857	29	1.983				
32	1.856	13	2.015				
33	2.060	33	2.060				
34	2.640	34	2.640				
35	2.769	35	2.769				

Note: The shaded items are the items deleted to form Test-24. In terms of difficulty, the order of items 1 – 24 remains the same in both Test-35 and Test-24.

Table 4:10

Distribution of Items in the Difficulty Group

	Test-35	Test-24
Easy items	1, 2, 5, 8, 9, 16, 18, 19, 20, 21, 22, 23 (12 numbers)	1, 2, 5, 8, 9, 16, 18, 19, 20, 21, 22, 23 (12 numbers)
Hard items	3, 4, 6, 7, 10, 11, 12, 13, 14, 15, 17, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35 (23 numbers)	3, 4, 6, 7, 10, 11, 12, 13, 14, 15, 17, 24 (12 numbers)

Table 4:11

Comparability of the Invariance of Examinee Ability Estimates in Different Item Groups

Item Groups	Correlation Coefficients	
	Test-35	Test-24
Equivalent-halves	0.580** (0.622**)	0.342** (0.554**)
Content	0.527** (0.573**)	0.285** (0.522**)
Difficulty	0.013* (0.477**)	0.023** (0.421**)

** Correlation is statistically significant at the 0.01 level (2-tailed).

* Correlation is statistically significant at the 0.05 level (2-tailed).

Note: The corresponding correlation coefficients after deleting the outliers are presented within parentheses

Table 4:12

Correlation Coefficient between Content-area-based and Total-test-based Difficulty Estimates within Individual Subject Groups

Group	Correlation coefficient between content-area-based and total-test-based estimates	
	Test-35	Test-24
Whole Group	1.000	1.000
Male Group	1.000	1.000
Female group	1.000	1.000
High-ability Group	1.000	0.992
Low-ability Group	0.982	0.989

Note: All correlations are statistically significant at the 0.01 level (2-tailed).

Table 4:13

Slope and Intercept of the Principal Axis for Each Subject Group

(Different Content Areas)

Group	Test-35		Test-24	
	Slope, β	Intercept, α	Slope, β	Intercept, α
Whole Group	1.034	-0.008	0.958	0.010
Male Group	1.020	-0.007	0.973	0.016
Female Group	1.047	-0.005	0.942	0.007
High-ability Group	0.884	0.012	1.024	0.059
Low-ability Group	1.259	-0.106	1.013	-0.028

Table 4:14

Mean Distances to Theoretical and Principal Axes

for Each Content Area

Group	Mean Distance	Content area	
		N&A	SSM
Whole Group	Distance from theoretical axis	0.0227 (0.0176)	0.0280 (0.0308)
	Distance from principal axis	0.0025 (0.0093)	0.0078 (0.0127)
Male Group	Distance from theoretical axis	0.0058 (0.0151)	0.0248 (0.0320)
	Distance from principal axis	0.0077 (0.0113)	0.0130 (0.0168)
Female Group	Distance from theoretical axis	0.0468 (0.0213)	0.0261 (0.0447)
	Distance from principal axis	0.0111 (0.0128)	0.0169 (0.0086)
High-ability Group	Distance from theoretical axis	0.1642 (0.1292)	0.1652 (0.1513)
	Distance from principal axis	0.0217 (0.1171)	0.0352 (0.1444)
Low-ability Group	Distance from theoretical axis	0.7069 (0.1056)	0.0969 (0.1771)
	Distance from principal axis	0.1575 (0.1003)	0.3355 (0.1791)

Note: The corresponding mean distances for Test-24 are presented within parentheses.

Table 4:15

Correlation Coefficient between Form-based

and Total-test-based Difficulty Estimates within Individual Subject Groups

Group	Correlation coefficient between form-based and total-test-based estimates	
	Test-35	Test-24
Whole Group	0.992	1.000
Male Group	0.993	1.000
Female group	0.988	0.999
High-ability Group	0.989	0.971
Low-ability Group	0.915	0.996

Note: All correlations are statistically significant at the 0.01 level (2-tailed).

Table 4:16

Slope and Intercept of the Principal Axis for Each Subject Group

(Different Forms)

Group	Test-35		Test-24	
	Slope, β	Intercept, α	Slope, β	Intercept, α
Whole Group	0.991	-0.012	0.931	-0.013
Male Group	0.979	-0.007	0.943	0.000
Female Group	1.002	-0.017	0.914	-0.027
High-ability Group	0.826	0.050	0.963	-0.011
Low-ability Group	1.253	-0.604	0.743	-0.041

Table 4:17

Mean Distances to Theoretical and Principal Axes for Each Item Form

Group	Mean Distance	Item Form	
		Regular	Comparative
Whole Group	Distance from theoretical axis	0.1121 (0.0553)	0.0799 (0.0191)
	Distance from principal axis	0.1023 (0.0104)	0.0884 (0.0269)
Male Group	Distance from theoretical axis	0.1007 (0.0431)	0.0663 (0.0099)
	Distance from principal axis	0.0841 (0.0047)	0.0791 (0.0243)
Female Group	Distance from theoretical axis	0.1353 (0.0734)	0.1061 (0.0452)
	Distance from principal axis	0.1330 (0.0168)	0.1068 (0.0334)
High-ability Group	Distance from theoretical axis	0.4146 (0.1985)	0.0832 (0.2653)
	Distance from principal axis	0.1580 (0.1610)	0.1762 (0.3210)
Low-ability Group	Distance from theoretical axis	0.5834 (0.4391)	0.5553 (0.4259)
	Distance from principal axis	0.6914 (0.0179)	0.2821 (0.0946)

Note: The corresponding mean distances for Test-24 are presented within parentheses.

Table 4:18

Point Biserial Correlations of the Items in Test-35 and Test-24

Item	Test-35					Test-24				
	Whole Group	Male Group	Female Group	High-ability Group	Low-ability Group	Whole Group	Male Group	Female Group	High-ability Group	Low-ability Group
1	0.353	0.340	0.335	0.133	0.114	0.391	0.376	0.382	0.170	0.232
2	0.198	0.204	0.203	0.044	0.157	0.226	0.230	0.231	0.068	0.229
3	0.287	0.278	0.292	0.148	0.101	0.333	0.319	0.343	0.224	0.202
4	0.360	0.383	0.333	0.224	0.085	0.401	0.423	0.377	0.304	0.172
5	0.354	0.364	0.341	0.161	0.140	0.402	0.410	0.393	0.236	0.251
6	0.370	0.408	0.335	0.224	0.101	0.424	0.458	0.394	0.318	0.216
7	0.259	0.272	0.238	0.271	0.085	0.278	0.290	0.258	0.318	0.125
8	0.361	0.372	0.347	0.139	0.197	0.402	0.410	0.391	0.179	0.306
9	0.359	0.367	0.344	0.130	0.164	0.406	0.408	0.399	0.188	0.270
10	0.290	0.298	0.281	0.181	0.121	0.336	0.339	0.332	0.280	0.196
11	0.286	0.282	0.293	0.264	0.118	0.304	0.298	0.311	0.316	0.145
12	0.315	0.310	0.315	0.248	0.100	0.344	0.339	0.344	0.297	0.150
13	0.292	0.319	0.246	0.324	0.061	0.315	0.349	0.264	0.385	0.094
14	0.229	0.250	0.193	0.199	0.074	0.247	0.270	0.211	0.246	0.103
15	0.330	0.343	0.308	0.270	0.078	0.353	0.367	0.332	0.344	0.097
16	0.364	0.364	0.355	0.177	0.227	0.385	0.384	0.377	0.216	0.245
17	0.337	0.357	0.313	0.163	0.188	0.366	0.383	0.346	0.236	0.207
18	0.425	0.448	0.402	0.173	0.242	0.443	0.469	0.417	0.223	0.239
19	0.455	0.452	0.458	0.140	0.254	0.474	0.473	0.474	0.187	0.259
20	0.488	0.492	0.471	0.154	0.229	0.509	0.512	0.495	0.220	0.218
21	0.481	0.479	0.468	0.171	0.210	0.485	0.489	0.468	0.224	0.159
22	0.361	0.366	0.357	0.104	0.233	0.354	0.363	0.344	0.143	0.168
23	0.394	0.407	0.379	0.108	0.250	0.383	0.401	0.364	0.131	0.187
24	0.288	0.291	0.282	0.221	0.181	0.259	0.275	0.239	0.229	0.089
25	0.353	0.359	0.343	0.251	0.173					
26	0.266	0.279	0.240	0.320	0.125					
27	0.496	0.514	0.458	0.275	0.138					
28	0.358	0.361	0.345	0.156	0.190					
29	0.291	0.297	0.275	0.313	0.130					
30	0.404	0.415	0.383	0.330	0.091					
31	0.274	0.269	0.264	0.107	0.205					
32	0.252	0.274	0.220	0.293	0.143					
33	0.226	0.231	0.216	0.334	0.137					
34	0.109	0.097	0.114	0.134	0.115					
35	0.166	0.188	0.129	0.275	0.097					
Mean	0.329	0.338	0.313	0.206	0.151	0.370	0.379	0.356	0.238	0.191
SD	0.099	0.101	0.097	0.081	0.058	0.085	0.086	0.086	0.078	0.063

Note: All the correlation coefficients were first transformed to Fisher's Zs. The means and standard deviations of the transformed values were then calculated in the usual way. They are then re-transformed back to the Pearson correlation coefficients.

Table 4:19

Six Hardest Items for Each Subject Group

Test	Group	Six hardest items					
Test-35	Whole Group	35 (2.769)	34 (2.640)	33 (2.060)	13 (2.015)	29 (1.983)	32 (1.856)
	Male Group	35 (2.506)	34 (2.492)	33 (1.918)	29 (1.768)	13 (1.720)	32 (1.652)
	Female Group	35 (3.091)	34 (2.833)	13 (2.355)	33 (2.235)	29 (2.232)	32 (2.091)
	High-ability Group	34 (3.336)	35 (3.034)	33 (1.894)	32 (1.422)	29 (1.296)	13 (1.173)
	Low-ability Group	35 (5.500)	30 (5.442)	34 (5.303)	29 (5.114)	33 (4.919)	13 (4.796)
Test-24	Whole Group	13 (1.879)	15 (1.604)	14 (1.576)	7 (0.976)	11 (0.803)	24 (0.801)
	Male Group	13 (1.598)	15 (1.401)	14 (1.355)	7 (0.820)	11 (0.736)	24 (0.696)
	Female Group	13 (2.190)	15 (1.824)	14 (1.815)	7 (1.139)	24 (0.912)	11 (0.878)
	High-ability Group	13 (1.023)	14 (0.962)	15 (0.492)	7 (0.078)	24 (-0.082)	11 (-0.268)
	Low-ability Group	13 (4.409)	15 (4.305)	14 (3.669)	24 (3.057)	7 (2.691)	6 (2.652)

Note: The numbers within parentheses are the item difficulty estimates.

Table 4:20

Percentages of Examinees Completing Test-35 and 75% of Test-35

Group	Percentages of Examinees	
	Whole Test	75% of the Test
Whole Group	43.0	76.2
Male Group	48.5	81.4
Female Group	38.3	71.6
High-ability Group	49.9	91.5
Low-ability Group	29.0	53.2

Table 4:21

Percentages of Examinees Omitting the Last 5 Items

Group	Percentages of Examinees					
	Item 31	Item 32	Item 33	Item 34	Item 35	All 5 items
Whole Group	30.4	33.0	34.5	37.6	37.4	27.9
Male Group	25.1	26.9	28.6	31.5	31.4	23.0
Female Group	35.0	38.3	39.7	42.9	42.5	32.1
High-ability Group	20.4	21.6	23.9	30.3	30.5	16.4
Low-ability Group	48.9	52.1	53.1	53.8	53.6	46.7

Table 4:22

Comparability of CTT- and Rasch-based Item Difficulty Estimates

Group	Correlation coefficient between CTT- and Rasch-based item difficulty estimates	
	Test-35	Test-24
Whole Group	1.000	1.000
Male Group	1.000	1.000
Female Group	1.000	1.000
High-ability Group	0.999	0.999
Low-ability Group	0.999	1.000

Note: All correlations are statistically significant at the 0.01 level (2-tailed).

Table 4:23

Comparability of Item Ordering based on CTT and Rasch Item Difficulty Estimates

(a) Test-35

Item Number									
Whole Group		Male Group		Female Group		High-ability Group		Low-ability Group	
CTT	Rasch	CTT	Rasch	CTT	Rasch	CTT	Rasch	CTT	Rasch
2	2	2	2	2	2	2	2	2	2
1	1	1	1	8	8	1	1	1	1
8	8	8	8	5	5	8	8	8	8
5	5	5	5	1	1	5	5	5	5
16	16	16	16	16	19	19	19	16	16
19	19	19	19	19	16	20	20	9	9
18	18	18	18	18	18	21	21	18	18
9	9	9	9	9	9	18	18	19	19
20	20	20	20	23	23	16	16	3	3
23	23	21	21	20	20	9	9	17	17
21	21	23	23	22	22	23	23	22	22
22	22	22	22	21	21	22	22	23	23
17	17	17	17	17	17	27	27	20	20
3	3	3	3	3	3	17	17	10	10
4	4	4	4	6	6	4	4	4	11
6	6	27	27	4	4	6	6	11	4
10	10	6	6	10	10	3	3	21	21
28	28	28	28	12	12	28	28	12	12
12	12	10	10	11	11	12	12	6	6
27	27	31	31	28	28	10	10	7	7
31	31	12	12	24	24	11	11	24	24
24	24	24	24	31	31	25	25	31	31
11	11	11	11	27	27	24	24	28	28
7	7	7	7	7	7	31	31	14	14
25	25	25	25	25	25	7	7	25	25
30	30	30	30	30	30	30	30	26	26
14	14	14	14	14	14	15	15	27	27
15	15	15	15	15	15	26	26	15	15
26	26	26	26	26	26	14	14	32	32
32	32	32	32	32	32	13	13	13	13
29	29	13	13	29	29	29	29	33	33
13	13	29	29	33	33	32	32	29	29
33	33	33	33	13	13	33	33	34	34
34	34	34	34	34	34	35	35	30	30
35	35	35	35	35	35	34	34	35	35

(b) Test-24

Item Number									
Whole Group		Male Group		Female Group		High-ability Group		Low-ability Group	
CTT	Rasch	CTT	Rasch	CTT	Rasch	CTT	Rasch	CTT	Rasch
2	2	2	2	2	2	2	2	2	2
1	1	1	1	8	8	1	1	1	1
8	8	8	8	5	5	8	8	8	8
5	5	5	5	1	1	5	5	5	5
16	16	16	16	16	19	19	19	16	16
19	19	19	19	19	16	20	20	9	9
18	18	18	18	18	18	21	21	18	18
9	9	9	9	9	9	18	18	19	19
20	20	20	20	23	23	16	16	3	3
23	23	21	21	20	20	9	9	17	17
21	21	23	23	22	22	23	23	22	22
22	22	22	22	21	21	22	22	23	23
17	17	17	17	17	17	17	17	20	20
3	3	3	3	3	3	4	4	10	10
4	4	4	4	6	6	6	6	4	11
6	6	6	6	4	4	3	3	11	4
10	10	10	10	10	10	12	12	21	21
12	12	12	12	12	12	10	10	12	12
24	24	24	24	11	11	11	11	6	6
11	11	11	11	24	24	24	24	7	7
7	7	7	7	7	7	7	7	24	24
14	14	14	14	14	14	15	15	14	14
15	15	15	15	15	15	14	14	15	15
13	13	13	13	13	13	13	13	13	13

Table 4:24

Comparability of CTT- and Rasch-based Examinee Ability Estimates

Group	Correlation coefficient between CTT- and Rasch-based ability estimates	
	Test-35	Test-24
Whole Group	0.960	0.911
Male Group	0.950	0.870
Female Group	0.969	0.951
High-ability Group	0.889	0.714
Low-ability Group	0.960	0.961

Note: All correlations are statistically significant at the 0.01 level (2-tailed).

Table 4:25

Comparability of Invariance of the CTT- and Rasch-based Item Difficulty Estimates

Sampling Plan	Average Correlation Coefficient			
	Test-35		Test-24	
	CTT	Rasch	CTT	Rasch
Random samples	0.996 (0.263)	0.996 (0.273)	0.995 (0.236)	0.995 (0.232)
Gender group sampling	0.983 (0.078)	0.982 (0.081)	0.980 (0.177)	0.976 (0.072)
Ability group sampling	0.881 (0.034)	0.894 (0.039)	0.848 (0.062)	0.883 (0.041)

Note: Standard deviations are presented within parentheses.

Table 4:26

Comparability of Invariance of the CTT- and Rasch-based Examinee Ability Estimates

Item Group	Correlation Coefficient			
	Test-35		Test-24	
	CTT	Rasch	CTT	Rasch
Equivalent-halves	0.644 (0.634)	0.580 (0.622)	0.589 (0.565)	0.342 (0.554)
Content	0.595 (0.581)	0.527 (0.573)	0.560 (0.531)	0.285 (0.522)
Difficulty	0.529 (0.474)	0.013 (0.477)	0.482 (0.420)	0.023 (0.421)

Note: The corresponding correlation coefficients after deleting the outliers are presented within parentheses

Table 4:27

Comparability of the Numbers of Unfitted Items

from CTT and Rasch Modeling

(a) Test-35

Sampling Plan	Sample	Number of Unfitted Items		No. of Common Unfitted Items	Percentage of Common Unfitted Items
		CTT	Rasch		
Random sampling	r_sample1	17	10	8	22.9
	r_sample2	19	12	9	25.7
	r_sample3	21	7	3	8.6
	r_sample4	20	12	7	20.0
	r_sample5	19	12	8	22.9
Gender group sampling	m_sample1	16	9	6	17.1
	m_sample2	17	11	5	14.3
	m_sample3	17	11	5	14.3
	m_sample4	17	15	8	22.9
	m_sample5	15	9	7	20.0
	f_sample1	19	8	4	11.4
	f_sample2	23	6	4	11.4
	f_sample3	18	9	5	14.3
	f_sample4	20	8	4	11.4
	f_sample5	21	8	6	17.1
Ability group sampling	h_sample1	33	4	4	11.4
	h_sample2	32	4	4	11.4
	h_sample3	34	5	5	14.3
	h_sample4	33	5	5	14.3
	h_sample5	30	7	7	20.0
	L_sample1	35	4	4	11.4
	L_sample2	35	1	1	2.9
	L_sample3	35	4	4	11.4
	L_sample4	35	4	4	11.4
	L_sample5	35	3	3	8.6

(b) Test-24

Sampling Plan	Sample	Number of Unfitted Items		No. of Common Unfitted Items	Percentage of Common Unfitted Items
		CTT	Rasch		
Random sampling	r_sample1	4	6	3	12.5
	r_sample2	5	6	2	8.3
	r_sample3	5	4	2	8.3
	r_sample4	6	6	4	16.7
	r_sample5	6	8	4	16.7
Gender group sampling	m_sample1	7	5	4	16.7
	m_sample2	4	5	1	4.2
	m_sample3	8	4	3	12.5
	m_sample4	4	7	3	12.5
	m_sample5	4	5	3	12.5
	f_sample1	7	3	2	8.3
	f_sample2	6	3	2	8.3
	f_sample3	4	6	3	12.5
	f_sample4	7	5	4	16.7
	f_sample5	6	4	3	12.5
Ability group sampling	h_sample1	18	1	1	4.2
	h_sample2	19	2	2	8.3
	h_sample3	19	4	4	16.7
	h_sample4	18	3	3	12.5
	h_sample5	21	1	1	4.2
	L_sample1	22	0	0	0.0
	L_sample2	20	0	0	0.0
	L_sample3	22	0	0	0.0
	L_sample4	23	3	2	8.3
	L_sample5	22	3	2	8.3

Reference

Acton, F.S. (1970). *Numerical Methods that Work.* U.S.A.: Harper & Row Publishers.

Anastasi, A. & Urbina, S. (1997). *Psychological Testing.* New Jersey: Prentice Hall.

Baker, F.B. (1964). An Intersection of Test Score Interpretation and Item Analysis. *Journal of Educational Measurement*, 1, 23-28.

Baker, F.B. (1965). Origins of the Item Parameters X_{50} and β as a Modern Item Analysis Technique. *Journal of Educational Measurement*, 2, 167-180.

Bartram, D. (1990). Reliability and Validity. In Beech, J.R. & Harding, L. (1990). *Testing People – A Practical Guide to Psychometrics.* England: The NFER – NELSON Publishing Company Ltd.

Bejar, I.I. (1980). A Procedure for Investigating the Unidimensionality of Achievement Tests Based on Item Parameter Estimates. *Journal of Educational Measurement*, 17, 283-296.

Birnbaum, A. (1968). Some Latent Trait Models and their Use in Inferring an Examinee's Ability. In Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores.* Reading, Mass.: Addison-Wesley Publishing.

Bond, T.G. & Fox, C.M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences.* U.S.A.: Lawrence Erlbaum Associates, Inc. Publishers.

Brown, F.G. (1976). *Principles of Educational and Psychological Testing.* U.S.A.: Holt, Rinehart and Winston.

Carmines, E.G. & Zeller, R.A. (1979). *Reliability and Validity Assessment.* Beverly Hills CA: Sage

Carver, R.P. (1978). The Case Against Statistical Significance Testing. *Harvard Educational Review*, Vol. 48, No. 3, 378-399.

Casella, G. & Berger, R.L. (1990). *Statistical Inference.* U.S.A.: Duxbury Press.

Cronbach, L.J. (1951). Coefficient alpha and the Internal Structure of Tests. *Psychometrika*, 16, 297-334.

DfEE & QCA (1999). *Mathematics – The National Curriculum for England Key Stages 1 – 4.* London

Divgi, D.R. (1986). Does Rasch Model Really Work for Multiple Choice Items? Not If You Look Closely. *Journal of Educational Measurement*, 23, 283-298.

Embretson, S.E. & Reise, S.P. (2000). *Item Response Theory for Psychologists.* New Jersey: Lawrence Erlbaum Associates, Inc. Publishers.

Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. *Journal of Educational Measurement*, 58(3), 357-381.

Fisher, W.P. (1994). The Rasch Debate: Validity and Revolution in Educational Measurement. In Wilson, M. (ed.) (1994). *Objective Measurement – Theory into Practice, Volume 2.* U.S.A.: Ablex Publishing Corporation.

Fitz-Gibbon, C.T. (1984). Meta-analysis: An Explication. *British Educational Journal*, Vol. 10, No. 2, 135-144.

Fitz-Gibbon, C.T. (1996). *Monitoring Education – Indicators, Quality and Effectiveness.* London: Cassell.

Fitz-Gibbon, C.T. (2000). *Value Added for those in Despair: Research Methods Matter.* The Vernon-Wall Lecture for the annual meeting of the Education Section of the British Psychological Society.

Glass, G.V., McGaw, B. & Smith, M.L. (1981). *Meta-Analysis in Social Research.* London: Sage Publications.

Goldman, S.H. & Raju, N.S. (1986). Recovery of One- and Two-Parameter Logistic Item Parameters: An Empirical Study. *Educational and Psychological Measurement*, 46, 11-21.

Goldstein, H. (1979). Consequences of Using the Rasch Model for Educational Assessment. *British Educational Research Journal*, 5(2), 211-220.

Goldstein, H. & Wood, R. (1989). Five Decades of Item Response Modeling. *British Journal of Mathematical and Statistical Psychology*, 42, 139-167.

Green, K.E. (1991). Reliability, Validity and Test Score Interpretation. In Green, K.E. (1991) (ed.). *Educational Testing – Issues and Applications.* U.S.A.: Garland Publishing Inc.

Green, S.B., Lissitz, R.W. & Mulaik, S.A. (1977). Limitation of Coefficient Alpha as an Index of Test Unidimensionality. *Educational and Psychological Measurement*, 37, 827-838.

Gustafsson, J.E. (1980). Testing and Obtaining Fit of Data to the Rasch Model. *British Journal of Mathematical and Statistical Psychology*, 38, 205-233.

Hambleton, R.K. (1995). Meeting the Measurement Challenges of the 1990s and Beyond New Assessment Models and Methods. In Oakland, T. & Hambleton, R.K. (eds.) (1995). *International Perspectives on Academic Assessment.* U.S.A.: Kluwer Academic Publishers.

Hambleton, R.K. & Rovinelli, R.J. (1986). Assessing the Dimensionality of a Set of Test Items. *Applied Psychological Measurement*, 10(3), 287-302.

Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory – Principles and Applications.* U.S.A.: Kluwer – Nijhoff Publishing.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* U.S.A.: Sage Publications

Hattie, J.A. (1985). Methodological Review: Assessing Unidimensionality of Tests and Item. *Applied Psychological Measurement*, 9, 139-164.

Hopkins, K.D., Hopkins, B.R. & Glass, G.V. (1996). *Basic Statistics for the Behavioral Sciences (Third Edition).* U.S.A.: Allyn and Bacon.

Hulin, C.L., Drasgow, F. & Parsons, C.K. (1983). *Item Response Theory – An Application to Psychological Measurement.* U.S.A.: Dow Jones-Irwin.

Kaiser, H.F. (1970). A Second Generation Little Jiffy. *Psychometrika*, 35, 401-415.

Kingston, N.M. & Dorans, N.J. (1985). The Analysis of Item-Ability Regressions: An Exploratory IRT Model Fit Tool. *Applied Psychological Measurement*, 9(3), 281-288.

Kline, P. (1990). How Tests are Constructed. In Beech, J.R. & Harding, L. (1990). *Testing People – A Practical Guide to Psychometrics.* U.K.: The NFER-Nelson Publishing Company Ltd.

Linacre, J.M. & Wright, B.D. (1989). The "Length" of a Logit. *Rasch Measurement Transactions,* 3(2), 51-62.

Linacre, J.M. & Wright, B.D. (1994). Chi-Square Fit Statistics. *Rasch Measurement Transactions,* 8(2), 360-361.

Lord, F.M. (1952). A Theory of Test Scores. *Psychometric Monograph.* No.7

Lord, F.M. (1953a). The Relation of Test Score to the Trait Underlying the Test. *Educational and Psychological Measurement*, 13, 517-548.

Lord, F.M. (1953b). An Application of Confidence Intervals and of Maximum Likelihood to the Estimation of an Examinee's Ability. *Psychometrika*, 18, 57-75.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* U.S.A.: Lawrence Erlbaum Associates, Inc., Publishers.

Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Lumsden, J. (1961). The Construction of Unidimensional Tests. *Psychological Bulletin*, 58, 122-131.

McDonald, R.P. (1981). The Dimensionality of Tests and Items. *British Journal of Mathematical and Statistical Psychology*, 34, 100-117.

McDonald, R.P. (1982). Linear versus Nonlinear Models in Item Response Theory. *Applied Psychological Measurement*, 6(4), 379-396.

McDonald, R.P. & Ahlawat, K.S. (1974). Difficulty Factors in Binary Data. *British Journal of Mathematical and Statistical Psychology*, 27, 82-99.

Mead, R. (1976). *Asessing the Fit of Data to the Rasch Model.* A paper presented at the annual meeting of AERA, San Francisco.

Meijer, R.R. & Sijtsma, K. (2001). Person Fit Statistics – What is Their Purpose? *Rasch Measurement Transactions*, 15(2), 823.

Meredith, W. (1970). Poisson Distributions of Error in Mental Test Theory. *British Journal of Mathematical and Statistical Psychology*, 24, 49-82.

Phillips, S.E. (1986). The Effects of the Deletion of Misfitting Persons on Vertical Equating via the Rasch Model. *Journal of Educational Measurement*, 23(2), 107-118.

Rasch, G. (1966). An Item Analysis which takes Individual Differences into Account. *British Journal of Mathematical and Statistical Psychology*, 19, 49-57.

Rasch, G. (1980) (Expanded ed.). *Probabilistic Models for Some Intelligence and Attainment Tests.* USA: University of Chicago. (Published originally in 1960 by the Danish Institute for Educational Research, Copenhagen).

Reckase, M.D. (1979). Unifactor Latent Trait Models Applied to Multi-Factor Tests: Results and Implications. *Journal of Educational Statistics*, 4, 207-230.

Roskam, E.E. (1997). Models for Speed and Time-Limit Tests. In Van der Linden, W.J. & Hambleton, R.K. (eds.). *Handbook of Modern Item Response Theory.* Springer.

Shavelson, R.J. & Webb, N.M. (1991). *Generalizability Theory – A Primer.* London: Sage Publications.

Shavelson, R.J., Webb, N.M. & Rowley, G.L. (1989). Generalizability Theory. *American Psychologist*, 44(6), 922-932.

Slinde, J.A. & Linn, R.L. (1979). The Rasch Model, Objective Measurement, Equating, and Robustness. *Applied Psychological Measurement*, 3, 437-452.

Stephenson, G. (1961). *Mathematical Methods for Science Students.* Longman Group Limited.

Thorndike, R.M. (1997). *Measurement and Evaluation in Psychology and Education (6th Edition).* U.S.A.: Prentice Hall.

Tucker, L.R. (1946). Maximum Validity of a Test with Equivalent Items. *Psychometrika*, 11, 1-13.

Van der Linden, W.J. & Hambleton, R.K. (1997). Item Response Theory: Brief History, Common Models, and Extensions. In Van der Linden, W.J. & Hambleton, R.K. (eds.) (1997). *Handbook of Modern Item Response Theory.* Springer.

Verhelst, N.D., Verstralen, H.H.F.M. & Jansen, M.G.H. (1997). *A Logistic Model for Time-Limit Tests.* In Van der Linden, W.J. & Hambleton, R.K. (eds.). *Handbook of Modern Item Response Theory.* Springer.

Wainer, H. & Wright, B.D. (1980). Robust Estimation of Ability in the Rasch Model. *Psychometrika*, 45, 373-391.

Willms, J.D. (1992). *Monitoring School Performance – A Guide for Educators.* London: The Falmer Press.

Wright, B.D. (1977). Solving Measurement Problems with the Rasch Model. *Journal of Educational Measurement*, 14, 97-116.

Wright, B.D. (1992). IRT in the 1990s: Which Models Work Best? *Rasch Measurement Transactions, Part 2,* 6(1), 193-208.

Wright, B.D. (1996). Negative Information. *Rasch Measurement Transactions,* 10(2), 489-508.

Wright, B.D. & Panchapakesan, N. (1969). A Procedure for Sample-free Item Analysis. *Educational and Psychological Measurement*, 29, 23-48.

Yen, W.M. (1981). Using Simulation Results to Choose a Latent Trait Model. *Applied Psychological Measurement*, 5, 245-262.