9-1-2011

Evaluating IRT- and CTT-based Methods of Estimating Classification Consistency and Accuracy Indices from Single Administrations

Nina Deng
ndeng@educ.umass.edu

EVALUATING IRT- AND CTT- BASED METHODS OF ESTIMATING CLASSIFICATION CONSISTENCY AND ACCURACY INDICES FROM SINGLE ADMINISTRATIONS

A Dissertation Presented

By

NINA DENG

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

September 2011

Education
Research and Evaluation Methods Program

EVALUATING IRT- AND CTT- BASED METHODS OF ESTIMATING CLASSIFICATION CONSISTENCY AND ACCURACY INDICES FROM SINGLE ADMINISTRATIONS

A Dissertation Presented

By

NINA DENG

Approved as to style and content by:

_____
Ronald K. Hambleton, Chair

_____
Craig S. Wells, Member

_____
Daeyoung Kim, Member

_____
Christine B. McCormick, Dean
School of Education

DEDICATION

To my parents, who always teach me to love learning.

To my husband, who always has love and belief in me.

ACKNOWLEDGMENTS

This dissertation could not have been written without the support and friendship found in the Research and Evaluation Method Program at the University of Massachusetts Amherst. My first debt of gratitude must go to my advisor, Professor Ronald Hambleton. It has been a great privilege to work with Ron for five years and I have consistently felt inspired by his wisdom, expertise, charisma, and humanity. I am indebted for his invaluable guidance and generous support through not only this dissertation but the whole journey of doctoral study. I learned a lot from him. Both his professionalism and personality had a great influence and will continue inspiring me throughout my life.

My committee members, Professor Craig Wells and Professor Daeyoung Kim, deserve a special note of gratitude for their support and helpful suggestions. Especially I am grateful to Craig for he has watched over me since my first days in the program and been consistently willing to offer assistance and advice.

I would like to give special thanks, beginning with Professor Stephen Sireci, for his guidance through my graduate studies, and his deep commitment to help and work with students. I want to thank Professor Lisa Keller for her lively lectures, and Professor April Zenisky for the opportunities of working with her. Also I want to thank Peg Louraine who has made my life at REMP a lot easier.

I could never forget my fellow students. The people I have met have become my

dear officemates, close friends, and helpful counselors. All of them have made REMP such a great place because I have been consistently surrounded by support and friendship.

Finally, I want to thank Dr. Skip Livingston for his thought-provoking comments on this study, and the Center for Advanced Studies in Measurement and Assessment at the University of Iowa, where the DC/DA softwares were offered free to the public which has saved tremendous time for me in the study. There is one more person I can hardly forgot to mention, Professor Feifei Ye, who introduced me to the field when I was in China, and without her I could have never known psychometrics.

ABSTRACT

EVALUATING IRT- AND CTT- BASED METHODS OF
ESTIMATING CLASSIFICATION CONSISTENCY AND ACCURACY INDICES
FROM SINGLE ADMINISTRATIONS

SEPTEMBER 2011

NINA DENG, B.A., SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

M.A., SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Ronald K. Hambleton

Three decision consistency and accuracy (DC/DA) methods, the Livingston and Lewis (LL) method, LEE method, and the Hambleton and Han (HH) method, were evaluated. The purposes of the study were: (1) to evaluate the accuracy and robustness of these methods, especially when their assumptions were not well satisfied, (2) to investigate the "true" DC/DA indices in various conditions, and (3) to assess the impact of choice of reliability estimate on the LL method.

Four simulation studies were conducted: Study 1 looked at various test lengths. Study 2 focused on local item dependency (LID). Study 3 checked the consequences of IRT model-data misfit and Study 4 checked the impact of using different scoring metrics. Finally, a real data study was conducted where no advantages were given to any models or assumptions.

The results showed that the factors of LID and model misfit had a negative impact on "true" DA index, and made all selected methods over-estimate DA index. On the

contrary, the DC estimates had minimal impacts from the above factors, although the LL method had poorer estimates in short tests and the LEE and HH methods were less robust to tests with a high level of LID.

Comparing the selected methods, the LEE and HH methods had nearly identical results across all conditions, while the HH method had more flexibility in complex scoring metrics. The LL method was found sensitive to the choice of test reliability estimate. The LL method with Cronbach's alpha consistently underestimated DC estimates while LL with stratified alpha functioned noticeably better with smaller bias and more robustness in various conditions.

Lastly it is hoped to make the software be available soon to permit the wider use of the HH method. The other methods in the study are already well supported by easy to use software.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1.1. Background

In many testing contexts, it is necessary to classify examinees into mutually exclusive performance categories based on a set of predetermined standards (e.g., state testing programs such as the MCAS). The standards are defined as a series of cut scores obtained from a standard setting process. The classification of performance provides an easy and convenient way to describe and to interpret examinees' performance in terms of proficiency levels, and is used a lot in both educational and licensure exams. The simplest example is the binary classification of mastery/non-mastery or pass/fail decision by applying one cut score. Multiple classifications classify examinees into more than two categories, for example, needs improvement, basic, proficient, and advanced.

These assessments with proficiency classifications often have high-stakes consequences, such as, graduation/license requirements and school accountability. The No Child Left Behind (NCLB) Act (2002) has required statewide standardized achievement tests to report examinees' performance in terms of ordered proficiency levels and so does the National Assessment of Educational Progress (NAEP) program, which resulted in a high demand of assessments reporting proficiency categories and, in turn had a great impact on students, teachers and schools (Li, 2006).

Along with the increased demands of assessments classifying examinees into

ordered proficiency categories, the classical approaches to reliability estimates may no longer serve the purpose quite well. People realize that the consistency of classifications rather than the consistency of test scores is of more concern. The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999, p.35) calls that "when a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure …" Two commonly used decision consistency and accuracy (DC/DA) indices, agreement index *P* (Hambleton & Novick, 1973) and coefficient *kappa* (Swaminathan, Hambleton, & Algina, 1974), were proposed. A number of procedures have been developed to estimate the indices based on a single administration since double test administrations are almost never practical to carry out.

Popular single administration methods include the procedures proposed by Huynh (1976), Subkoviak (1976), Hanson and Brennan (1990), and Livingston and Lewis (1995). However, the above methods were all developed in the framework of classical test theory (CTT) and most were based on the assumption that the items are dichotomously scored and equally weighted (except for, Livingston & Lewis, 1995).

Lord and Novick (1968) were among the first to introduce the model-based measurement and started a quiet but profound revolution in test theory and practices. Item response theory (IRT) has become the mainstream in the current educational measurement field and is widely used in standardized tests in many aspects such as test

development, item calibration, test scoring, equating, standard setting, etc.

1.2. Statement of the Problem

IRT is a powerful technique that an increasing number of test developers are employing in various aspects of test development and analyses (for applications of IRT, see Hambleton, Swaminathan, & Rogers, 1991). Nevertheless, many current popular decision consistency and accuracy methods (e.g., the work by Huynh, Subkoviak, Livingston and Lewis) were developed in the framework of CTT. These methods, particularly the most popular method developed by Livingston and Lewis (1995), are widely used. It is not uncommon in many testing programs to observe that all test analyses are carried out in an IRT framework but the classification consistency/accuracy indices are calculated in the framework of CTT. This inconsistency justifies a further investigation of the performance of CTT-based methods for the data fitting IRT models. Besides, some IRT-based methods were developed more recently (see the work by Lee, Rudner, and Hambleton and Han). These methods are new and deserve further study. Lastly, all the methods were built upon certain assumptions and therefore it is of great interest to check their robustness to the conditions where their assumptions are not well met.

1.3. Purpose of the Study

The purpose of the study was to investigate the performance of one CTT-based method, the LL method, and two IRT-based methods, the LEE, and HH methods, in estimating classification consistency and accuracy indices in various test conditions

through a series of simulation studies and a real data study. The simulation studies were used as the main study because various test conditions could be conveniently created and the "true" values of DC/DA indices were known. The real data study was carried out as a supplemental approach in which the methods were evaluated using the two forms of a real test, where no advantages were given to any models or assumptions. Specifically, the following research questions were addressed in this study:

(1) How accurate are the Livingston and Lewis (LL) method, Lee (LEE) and Hambleton and Han (HH) methods with simulated data in the IRT framework? And, how do they function with real data?

(2) How robust are the selected methods to non-standard conditions, including short test lengths, local item dependence, IRT model misfit, and composite scoring? What are the "true" DC/DA indices in those non-standard conditions?

(3) Since the LL method is sensitive to reliability estimates, what is the impact if alternative choices of reliability estimates are used in the LL method?

1.4. Organization of the Thesis

This thesis begins with the problem and purposes of the study, followed by Chapter 2, which provides an introduction of the DC/DA concepts, and a comprehensive review of CTT- and IRT-based methods of estimating DC/DA in the literature, including the models, assumptions, and a detailed review for each of the major methods. Then a series of simulation studies are presented in Chapter 3, which consist of four independent studies, investigating the robustness of the selected methods in conditions of (1) various

test lengths, (2) local item dependency, (3) IRT model misfit, and (4) different scoring metrics, separately. Lastly, the selected methods are evaluated using real data in Chapter 4. This thesis concludes with a summary of the results, and a discussion of their implications for researchers and practitioners.

CHAPTER 2

LITERATURE REVIEW

This chapter begins with a description of classical reliability coefficient estimates and limitations, followed by an introduction of the concept and indices for decision consistency and decision accuracy, denoted as DC and DA, respectively. Next, a review of current CTT- and IRT-based methods of calculating DC and DA indices is provided, including models, assumptions, procedures, and relevant research. The chapter concludes with a discussion about the DC/DA statistics.

2.1. Classical Reliability Estimates

2.1.1. Definition of Classical Test Theory and Reliability Coefficient

When a test is administrated, it is for certain that test users want the test results to be replicated if the test were given to the same group of individuals repeatedly (with little change in true scores between). The desired property of consistent test scores is referred as reliability. The concept of reliability begins with the concern of precision of a measurement, which is not a sufficient but a necessary condition for a test to be valid. Strictly speaking, no test is completely free of errors. The observed scores from the repeated administrations of the same test won't be identical. However, the less the variance of these scores are, the more confidence we have with the scores. On the other side, if the observed scores fluctuate greatly from one administration to another, the validity of the test scores problematic. The consistency of results is desired for physical measurement too. If a box is weighted repeatedly and the scale reads quite different

numbers each time, obviously the scale is not accurate and you won't want to rely on it. Likewise, it is desirable to know whether a test could produce comparable scores in its repeated administrations, so that the precision of the scores as well as their usage can be fairly justified.

Unlike the measures in the physical world, the tests measuring people's mental abilities cannot be administrated to the same individuals again and again. The test scores won't keep the same due to some reasons such as memorization, practice effect, shift of ability, etc., even though the test itself is constructed satisfactorily reliable. Thus the reliability of a mental test needs to be estimated indirectly.

The classical test theory initiated by Charles Spearman is one of the most significant inventions and provides theory and statistical model for estimating test reliability. It begins with the assumption that an observed score on a test (X) may be modeled as the sum of the examinee's "true score" (T) and an error component (E), expressed as $X = T + E$. The examinee's true score can be interpreted as the average of observed scores if the test could be administrated to the examinee for an infinite number of times. The error component is specific to the particular observed score in the realized administration, which makes it different from the examinee's true score. Given the definitions of true and error scores, the *reliability coefficient* $\rho_{XX'}$, which is defined as the correlation between scores on repeated administrations of a test or parallel tests, can be mathematically expressed as the ratio of true score variance to the observed score variance

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}$$

Therefore $\rho_{XX'}$ can be interpreted as the proportion of observed score variance that can be attributed to the true score variance.

Besides for the correlation between observed scores on repeated administration tests, another important question is, what is the relationship between the examinee's observed score and true score? The *reliability index* ρ_{XT} is defined as the correlation coefficient between the observed and true scores of a test. Mathematically, $\rho_{XT} = \frac{\sigma_T}{\sigma_X} = \sqrt{\rho_{XX'}}$, and ρ_{XT} sets the upper limit for test validity, $\rho_{XY'}$. Therefore the test validity would be questionable if the reliability coefficient $\rho_{XX'}$ is low, since then the reliability index ρ_{XT} cannot be high.

2.1.2 Methods of Calculating the Reliability Coefficient

Given the definition that the reliability coefficient is the correlation between scores of repeated administrations of a test or two parallel tests, a straightforward way to calculate the reliability is to have a single group of examinees taking the same form twice, or taking two parallel forms of a test, and to calculate the Pearson product moment correlation coefficient. However, this approach is not usually realistic and could hardly give an accurate estimate either. First, strictly parallel forms are rare in reality. Secondly, the time period between test-retest plays a crucial role and affects the reliability estimates.

To overcome the drawbacks of the two-administration procedure, a

single-administration reliability estimate was proposed and has been widely applied. By using a single-administration approach, only one test administration to a single group of examinees is required. Different from the earlier definition, the reliability coefficient here is essentially the internal consistency coefficient, because we would calculate the correlation between separately scored parts of the single test. It is claimed that the specific items in the test is only part of a larger content domain, and it is reasonable to assume that the generalization of examinee's performance on the specific items to the larger content domain can be estimated by evaluating how consistently the examinee perform across different items in the single form. The single-administration reliability estimates have been used widely and many are reported as routinely today. Below are the descriptions of several popular methods.

Split-half methods require that the single test needs to be divided into halves before reliability is estimated. Two estimates are available based on the split-half procedure. The first one is called corrected split-half reliability estimate, which applies the Spearman-Brown formula to obtain the corrected estimate of the reliability coefficient for the original full-length test based on correlation coefficient between the two half-tests. The second one called split-half reliability estimate, also called Guttman's/Rulon's formula, uses the scores of the two half-tests to calculate the reliability estimate for the original full-length test, without applying the Spearman-Brown formula. The latter one is easy to calculate, and provides a lower bound estimate for reliability (equal to reliability when the two half-tests are parallel).

The shortcoming of the split-half procedure is that there are numerous ways to split the test, therefore, it cannot provide a unique estimate for the reliability coefficient.

In the 1930s and 1940s, a second class of methods analyzing the variance-covariance structure of the item responses emerged and was rapidly developed. The coefficient alpha, also known as Cronbach's alpha, for polytomous items and Kuder Richardson 20 for binary items are most well-known and widely used today. The coefficient alpha can be used in any situation where the reliability of a composite is estimated. Most commonly it treats each item as a component of the test. If it treats the test consist of two half-tests, it is identical to the Guttman's/Rulon's formula. Cronbach (1951) illustrated the relationship between the coefficient alpha and split-half estimates: coefficient alpha is the average of all possible split-half estimates using Guttman's/Rulon's formula. It is worthy to point out that the coefficient alpha is based on the assumption of all the components in the test being perfectly parallel, which is unlikely to happen. Therefore coefficient alpha provides a lower bound of the reliability coefficient rather than a direct estimate. For example, if an alpha value of 0.8 is obtained, it is safe to say that at least 80% of the observed score variance is due to true score variance.

Kuder Richardson 20 (KR-20) is a special case of coefficient alpha when all the items are dichotomously scored. KR-21 was derived assuming that all the binary items are equally difficult. Therefore KR-21 is in turn a special case of KR-20. KR-21 is systematically lower than KR-20 and gives a lower bound and a quick estimate of

KR-20, when the more complicated computation is unlikely to be available.

As discussed above, coefficient alpha provides an accurate estimate for reliability when all the items in the test are perfectly parallel, which, however, is rare. In most cases, coefficient alpha provides a lower bound for reliability and underestimates the reliability coefficient. Stratified coefficient alpha was thus proposed as a more appropriate estimate by treating items in different content or cognitive categories as separate subtests (Rajaratnam, Cronbach, & Gleser, 1965) when calculating the reliability estimate. It is argued that when the test consists of items from distinct content categories, stratified alpha provides a substantially more accurate estimate of reliability (Cronbach et al. 1965) and nearly always higher in value.

2.1.3. Limitations of Classical Reliability Estimates

Glaser (1963) pointed out that the scores on an achievement test could provide two kinds of information. One kind is the relative position of the examinee's score in terms of the score distribution. The second kind is the degree to which the examinee has mastered the goals of instruction. The tests can be categorized as norm-referenced tests (NRT) or criterion-referenced tests (CRT) based on how the scores are interpreted. The different purposes and usage of the tests also determine how the test scores are reported, including raw score, scaled score, percentile, proportion of correct answers, etc. One of the most popular ways in reporting is to classify the examinees into multiple mastery levels. The classification of proficiency levels is widely used in CRT in which the examinee's proficiency level is determined by applying cut scores in relation to a

well-defined domain of knowledge and skills, which is usually derived from a standard setting procedure. The examinees are usually classified into two (passing and failing) or more proficiency levels (e.g., failing, basic, proficient, and advanced). Classification of proficiency is widely used in many testing programs such as state achievement tests and credentialing exams. The classification can also be used in NRT if the decision about examinees is made in terms of their position in the score distribution; however, the application is rare.

With the increasing use of proficiency classification, test users may be concerned with some questions such as, what is the expected proportion of examinees who would be classified consistently upon retesting? What is the probability that an examinee with true score above a cut score would be classified as a non-master? The accuracy and consistency of classifications, rather than the scores, become the central concern in such circumstances. This concern becomes even more compelling with more consequences associated with the decision made in terms of examinees' proficiency levels. For example, the decision may be used (1) to evaluate teachers and schools' performance, (2) to determine students' ability to graduate, or (3) to decide whether a certificate is issued or not. The classical reliability estimate, which was developed based on continuous test scores, is no longer appropriate to assess the classification consistency. New techniques for assessing reliability are needed.

2.2. Concepts and Indices of DC/DA

As has been discussed, the accuracy and consistency of classifications are of the

most interest when the tests are used to make classifications about examinee performance. The concepts of DC and DA were proposed as the indices to describe the reliability and validity of classifications (see Hambleton & Novick, 1973). Decision consistency refers to, when the test is used to make categorical decisions, the extent to which the classifications agree based on two independent administrations of the test (or two parallel forms of the test). Decision accuracy refers to the extent to which the actual classifications based on observed scores agree with the "true" classification based on true scores. Analogously, decision consistency concerns the reliability of the classifications, while decision accuracy concerns the validity of the classifications. It is worthy to point out that the value of DA is higher than that of DC. This is the case because the calculation of DC involves two sets of observed scores, while in calculating DA, only one set of observed scores is involved, the other set is true scores, which, are free of measurement error.

2.2.1. Agreement Indices P and Kappa

Hambleton and Novick (1973) proposed the agreement index *P* as a measure of decision consistency. This notion not only underlies the concept of DC, but also introduced a large body of literature since then devoted to formulation and estimation of reliability coefficient for proficiency classifications. Agreement index *P* is defined as the proportion of examinees consistently classified on alternative administrations of a test. It can be expressed as

$$P = \sum_{j=1}^{J} p_{jj}$$

where p_{jj} is the proportion of examinees consistently classified into the j^{th} category across the two administrations of the test, and *J* is the number of performance categories. For example, suppose a single cut score divides the examinees into passing and failing groups (J=2), and the rows and columns in Figure 2.1 represent the two administrations of the same test. Let p_{00} represent the proportion of examinees classified as failing in both measures, and p_{11} the proportion of examinees classified as passing in both measures. The index of decision consistency is $P = p_{00} + p_{11}$. If Administration 1 in Figure 1 is replaced with one set of observed scores, and Administration 2 is replaced with the true scores (or another criterion measure), *P* then becomes the decision accuracy index. In addition, p_{10} represents the proportion of examinees who are true masters but classified as failing, and p_{01} represents the proportion of examinees who are true non-masters but classified as passing. It is common for p_{10} to be termed as the false negative error rate, and p_{01} termed as the false positive error rate. Both indices reflect the classification inconsistency and are commonly reported in the evaluation of decision accuracy. Based on the purposes and uses of specific tests, one index is often of more concern than the other.

Some suggestions have been made to transform *P* to a more interpretable measure of decision consistency, or at least a measure that is less influenced by chance agreement. One of the most popular ones was made by Swaminathan, Hambleton, and

Algina (1974), who suggested the use of Cohen's *kappa* (Cohen, 1960) to correct for chance consistency,

$$k = \frac{P - p_c}{1 - p_c}, \text{ where}$$

$$p_c = \sum_{j=1}^{J} p_{j.} p_{.j}$$

where k is the kappa coefficient, p_c is chance agreement, J is the number of categories, and $p_{j.}$ and $p_{.j}$ are the marginal proportions of examinees falling in the j^{th} category in the two administrations, respectively. p_c stands for the decision consistency expected by chance, that is, when the two administrations are statistically independent. And *kappa* measures the test's contribution to the overall decision consistency beyond which is expected by chance. k has a value between 0 and 1. A value of 0 means that the decisions are as consistent as the decisions based on two tests which are statistically independent; a value of 1 means that the decisions are as consistent as the decisions based on two tests which have perfect agreement. Later Agresti (2002) describes a refinement of P in which larger discrepancies between the two administrations indicate more lack of agreement.

Someone argued that p_c was actually the proportion of consistency expected from the group consisting of particular marginal frequencies (Subkoviak, 1980). It is therefore suggested to report *kappa* together with the information of the particular testing situation, including the marginal proportions, test length, score variability, location of cut scores, etc.

2.2.2. Alternative Agreement Indices

The above described agreement indices *P* and *kappa* that reflect the decision consistency while treating all the false classifications equally seriously. That is, the misclassifications which are far above or below the cut score are not treated more serious than the ones which are near the cut score. The type of indices was referred as threshold loss agreement indices in the literature. Alternatively, some coefficients have been developed to reflect the various degrees of misclassifications. The second type was referred as squared-error loss agreement indices in the literature (Berk, 1980; Traub & Rowley, 1980). Major coefficients of the second kind include Livingston's k^2 and Brennan and Kane's $\Phi(\lambda)$. Both indices formulate the decision consistency depending on two factors: the test score generalizability and the difference between the mean score and the cut score.

Livingston (1972) defined k^2 as:

$$k^2 = \frac{\sigma_T^2 + (\mu_X - C)^2}{\sigma_T^2 + (\mu_X - C)^2 + \sigma_E^2}$$

where *C* is the cut score, μ_X is the mean test score, σ_T^2 is the variance of true scores, and σ_E^2 is the error variance. If $C = \mu_X$, k^2 is essentially the classical test theory reliability coefficient. Therefore k^2 is a generalization of the classic reliability coefficient.

Brennan and Kane (1977) derived the dependability index $\Phi(\lambda)$ in the framework of generalizability theory to represent decision consistency. The index is

defined as follow:

$$\Phi(\lambda) = \frac{\sigma_P^2 + (\mu - \lambda)^2}{\sigma_P^2 + (\mu - \lambda)^2 + (\sigma_i^2 + \sigma_E^2)}$$

where σ_P^2 is variance of persons, μ is the grand mean over persons and items, λ is the cut score on the percent score scale, and σ_E^2 is the error variance. In addition, it introduces a component of variance due to item difficulty σ_i^2. It has been shown that if all the items are dichotomously scored and $\lambda = \mu$, $\Phi(\lambda)$ reduces to the KR-21 coefficient. Since $\Phi(\lambda)$ introduces the variance of item difficulties in the denominator, k^2 is always larger than $\Phi(\lambda)$. Both indices have the advantages of providing more information with regard to the magnitude of misclassification, which is helpful if the test users want to know more in addition to the decision consistency proportion.

2.2.3. Factors Affecting P and Kappa

The agreement indices of *P* and *kappa* are easy to understand and interpret and are widely reported as the decision consistency index. The factors affecting *P* and *kappa* were of wide interest and extensively studied in the literature. Previous studies showed that *P* and *kappa* might be affected by the factors including the location of cut scores, test length, score variability, test score distribution, and the classical test score reliability. However, the sensitivity of *P* and *kappa* to these factors may not display in the same way.

Many studies showed that an increased test length, an increased classical test score reliability, or an increased score variability, keeping other conditions unchanged,

resulted in higher values of both *P* and *kappa* (Crocker & Algina, 1986; Huynh, 1976; Berk, 1980). The influence of the location of cut score, however, seemed more complicated. When the cut score moved away from the center of the score distribution, *P* increased while *kappa* decreased. Possible explanation was suggested by Huynh (1976) that since the chance agreement is near 1 when the cut score is at the tails of a score distribution, there seems to be not much room for improvement of the decision consistency beyond the chance consistency. As a result, the value of *kappa* dropped when the cut score became too small or too large. On the contrary, *P* became the largest when the cut score was away from the center. Since there were fewer examinees near the cut score, misclassification was less likely to happen then. However, a large proportion of the increased *P* associated with far-away-from-center cut score is due to the increased chance agreement. In addition, it was found that more cut score points would result in a lower value of *P* since this would result in more candidates being close to cut score points and with an increased chance of being misclassified.

There have been some discussions in the literature about whether *P* or *kappa* is a more appropriate index. However, no explicit conclusion can to be drawn. It is not suggested favoring one index over the other. Rather, they are alternative ways in estimating decision consistency as long as their interpretations are correctly understood. Nonetheless, comparisons between the two indices have been made. Wan, Brennan, and Lee (2007) found that *kappa* was more sensitive to the magnitude of reliability estimates than *P*, and higher *P* not always associated with higher *kappa*. Besides, *kappa*

was criticized due to its assumption of exact marginal proportions (Berk, 1980). An example made by Livingston and Wingersky (1979) illustrated that if 87% of the examinees passed the exam, kappa will correct the chance agreement based on the assumption that "chance agreement" would result in exactly 87% of examinees passing the exam. It is argued by the above researchers that *P* is more useful for tests where an absolute cut score is chosen, while *kappa* makes more sense when the cut score is determined by the consequences of the passing/failing proportion.

2.3. Methods of Estimating P and Kappa

The notions of agreement index *P* and its corrected form *kappa* proposed by Hambleton and Novick (1973) and Swaminathan, Hambleton, and Algina (1974) not only conceptualized decision consistency and accuracy but also realized the procedures for estimating the indices. More importantly, they initialized the practice of reporting *P* and *kappa* for tests used to make mastery classifications which is widely done today. The *Standards* (AERA, APA, & NCME, 1999, p.35) in its most recent version call for estimates of proportion of examinees who would be consistently classified using the same or alternative forms whenever a test is used to make categorical decisions. The procedure of calculating *P* described by Hambleton and Novick (1973) is quite straightforward and was deemed as the easiest method to understand, compute and interpret (Berk, 1980). Nevertheless, it is obvious that two administrations of a test are required if this procedure is adopted, which is often unrealistic and inconvenient in practice.

The single-administration approach for estimating *P* and *kappa* was introduced and developed to overcome the restriction of the two-administration procedure, led by Huynh (1976), Subkoviak (1976) and other researchers. Analogous to the split-half reliability coefficient estimate, researchers tried splitting a test into halves for estimating the DC index, however, no "step-up" formula for a corrected estimate generalizing from a half-test estimate to full-length test was available, not to mention the problem that there is non-uniqueness in splitting the test into halves. Alternatively, new models were introduced which helped the advancement of estimating *P* and *kappa* using a single administration. The single-administration methods are discussed in details in this chapter.

2.3.1. Models for Estimating P and Kappa

The role of the models in estimating classification indices is to estimate the true score distribution and to predict the observed score distribution of an alternative administrations of the test conditional on true score level. By assuming certain measurement models for the test data, the single-administration methods estimated the true score and conditional observed score distributions, then the J x J classification contingency tables can be constructed, and the agreement index *P* and *kappa* can be computed based on the tables. The parameters of the models, distributions of true and observed scores, and in turn the classification indices are all estimated based on the actual data from a single test administration (Lee, Hanson, & Brennan, 2002). Below are the descriptions of popular measurement models assumed in single-administration

methods.

Binominal model, one of several strong true-score models, was typically used to predict the probability of getting test score x given true ability π. The probability can be expressed as

$$P(x \mid \pi) = \binom{n}{x} \pi^{x} (1-\pi)^{(n-x)}, \; x = 0, 1, \ldots, n.$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

where n is the number of items, x is test score (number of correct answers) ranging from 0 to n; π is defined as the proportion of items out of all the items in the domain that the examinee can answer correctly, it is therefore on the percent-correct scale and called domain score or relative true score; n is the total number of items.

Under the 2-parameter beta binomial model (2PB), the conditional distribution of x given π is assumed to be binomially distributed, in addition, the density of π is assumed to be a beta distribution with two shape parameters, α and β (Keats & Lord, 1962). Thus the density of test score x for n items becomes (Huynh, 1976)

$$f(x) = \binom{n}{x} B(\alpha + x, n + \beta - x) / B(\alpha, \beta)$$

where B is the beta function with parameter α and β, x is test score, and n is the total number of items. The parameters α and β can be estimated using KR-21 and the first two moments (mean and standard deviation) of the observed score distribution.

The 4-paramter beta binomial model (4PB) assumes the true proportion-correct

score π with a beta distribution with four parameters: α, β, and additionally, the lower and upper limits, *a* and *b* (Lord, 1965). Different from $0 \le \pi \le 1$ in the 2PB model, π is set as $a \le \pi \le b$ under the 4PB model. The 4PB model was proven to have better performance than the 2PB model in fitting the observed score distributions (Hanson & Brennan, 1990).

The binomial model is built assuming that all the items are independent and equally difficult. However, studies showed that the violation of the assumption of equal difficulty did not very much affect the results (Subkoviak, 1978; Spray & Welch, 1990).

Multinomial model was introduced to estimate the probability of getting a summed score given the true ability for tests consisting of polytomously scored items which have the same number of categories (Lee, 2007; Lee, Brennan, & Wan, 2009). For example, a test consists of *n* polytomous items which have the same number of score categories, say, *k* categories. It assumes that the true abilities required for getting *k* possible item values, denoted as $\pi_1, \pi_2, ..., \pi_k$, are the same across the items. Following a multinomial model, the probability for an examinee with true abilities $\pi_1, \pi_2, ..., \pi_k$ getting x_1 items scored c_1, x_2 items scored c_2,…, and x_k items scored c_k, can be expressed as

$$P(x_1, x_2, ..., x_k \mid \pi_1, \pi_2, ..., \pi_k) = \frac{n!}{x_1!x_2!...x_k!}\pi_1^{x_1}\pi_2^{x_2}...\pi_k^{x_k}$$

where $\pi_1, \pi_2, ..., \pi_k$ denote the true abilities required to get the *k* possible item values, $c_1, c_2, ..., c_k$ are the *k* possible item values, and $x_1, x_2, ..., x_k$ are the observed

numbers of items getting each of the k possible values. Note that $\sum_{j=1}^{k} x_j = n$, and $\sum_{j=1}^{k} \pi_j = 1$. Thus the probability for an examinee with true abilities $\pi_1, \pi_2, ..., \pi_k$ getting a summed score X is

$$P(X \mid \pi_1, \pi_2, ..., \pi_k) = \sum_{c_1 x_1 + c_2 x_2 + ... + c_k x_k = X} P(x_1, x_2, ..., x_k \mid \pi_1, \pi_2, ..., \pi_k)$$

Note that the multinomial model reduces to the binomial model when all the items are dichotomously scored.

Compound multinomial (CM) model is used when the test is a mixture of dichotomous and polytomous items, or consists of polytomous items that differ in terms of the number of score categories, or both. Under the CM model, the items with the same number of categories are viewed as an item set. The probability of a summed score y for item set i is denoted as $P(y_i \mid \vec{\pi}_i)$, where $\vec{\pi}_i$ is the true ability and expressed as $\{ \pi_{1i}, \pi_{2i}, ..., \pi_{ki} \}$. The probability of a vector of summed scores for L item sets is

$$P(y_1, y_2, ..., y_L \mid \vec{\pi}_1, \vec{\pi}_2, ..., \vec{\pi}_L) = \prod_{i=1}^{L} P(y_i \mid \vec{\pi}_i)$$

And the probability of the total summed score z for the test is

$$P(z \mid \vec{\pi}_1, \vec{\pi}_2, ..., \vec{\pi}_L) = \sum_{y_1, ..., y_L : \sum w_i y_i = z} P(y_1, y_2, ..., y_L \mid \vec{\pi}_1, \vec{\pi}_2, ..., \vec{\pi}_L)$$

where w_i is the weight of the summed score of item set i. Note that when L =1, the CM model reduces to the multinomial model.

Item response theory (IRT) was introduced most prominently in Lord and Novick

(1968) and has been increasingly used in many aspects of test development and analyses. The IRT models assume that there is a latent trait θ underlying all the item responses and the item responses are independent after θ is controlled for. Using IRT, the relationship between examinee's latent ability θ and the responses to item i U_i can be modeled using a family of logistic models. The popular IRT models in the family include the 1-parameter, 2-parameter, and 3-parameter logistic models for dichotomous items, and the graded response model, and the generalized partial credit model for polytomous items. The mathematical expression of the three-parameter logistic (3PL) IRT model is shown below as an example.

$$P(U_i = 1 | \theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7 a_i (\theta - b_i)]}$$

where $P(U_i = 1 | \theta)$ is the probability of having response of 1 (correct answer) to item i given latent ability θ. a_i, b_i and c_i are item discriminating parameter, item difficulty parameter, and item guessing parameter, respectively. 1.7 is the scaling factor introduced to approximate a two-parameter normal ogive function with the same values for the a and b item parameters in the logistic model.

2.3.2. Assumptions of Methods

The current methods for single administration estimates of decision consistency and decision accuracy can fall into two general categories in terms of the psychometric foundations upon which the methods were built: the CTT-based approach and the IRT-based approach.

Early methods were developed under the CTT framework. The binomial model, also classified as a strong true-score model because of the assumptions made that go beyond those of the CTT model, was assumed for the observed score distributions of tests which consist of dichotomous items only. Popular methods were developed by Huynh (1976), Subkoviak (1976), and Hanson and Brennan (1990). Later, Livingston and Lewis (1995) extended the binomial model to handle polytomous items. Alternatively, Lee and his colleagues (Lee, 2007; Lee, Brennan & Wan, 2009) introduced multinomial and compound multinomial models for polytomously-scored items. In addition to the analytical approach, Brennan and Wan (2004) developed a bootstrap procedure for complex assessment based on the binomial and multinomial models.

The above methods can further be divided to two types concerning the assumption made for true score distributions: the *distributional* approach and the *individual* approach (Brennan & Wan, 2004; Lee, 2005). The distributional approach makes a distributional assumption for true abilities, e.g., Huynh (1976), Hanson and Brennan (1990). Livingston and Lewis (1995) assumed a family of beta distributions for the true score distributions. On the contrary, the individual approach calculates the decision consistency index for each examinee at one time and averages across all examinees, without making any assumption about true ability. Examples of the second type include Subkoviak (1976), Lee (2007), Lee, Brennan & Wan (2009), and Brennan and Wan (2004).

Due to the complexity in calculating the beta-binomial distribution, some researchers have proposed a normal approximation by assuming a bivariate normal distribution for the observed score distributions across two test administrations, with a correlation equal to test score reliability. Several methods exist, however, they differ in the way of calculating the reliability. Peng and Subkoviak (1980) used the KR-21 coefficient as test reliability. Woodruff and Sawyer (1989) split the original test into halves and applied the Spearman-Brown formula to get an estimate of the test reliability. Breyer and Lewis (1994) also adopted the split-half approach but employed a separate cut score for each of the two half-tests, and used a tetrachoric correlation in calculating the reliability.

The IRT-based approach has been developed along with an increasing popularity of IRT applications in various aspects of testing practice. Essentially, all the IRT- based methods were developed based on the same assumptions as are made with other IRT applications, including unidimensionality, local item independency and model fit. In addition, large sample sizes are needed for accurate estimation of item parameters. At the same time, it is not known how consequential random errors in the item parameter estimates due to small sample size might be on the stability and accuracy of single administration estimates of DC and DA.

Under the IRT framework, several methods were developed for tests scored on the raw test score scale. Earlier studies included Huynh (1990) using the Rasch model, and Wang, Kolen, and Harris (2000) using polytomous IRT models. More recently, Lee

(2010) developed a procedure which can be used for a mixture of IRT models. Briefly, these methods used IRT models to calculate the probability of a vector of responses conditional on latent ability θ, and then employed the compound binomial/multinomial model to calculate the observed raw score distribution conditional on θ. The raw score distribution integrated over θ can be achieved finally either by assuming a distribution for θ or by using individual θ estimates.

From a different perspective, Rudner (2001, 2005) developed a procedure for tests scored on the θ scale. It assumed that the conditional distribution of estimated ability $\hat{\theta}$ followed a normal distribution with a mean of θ and standard deviation of SE($\hat{\theta}$). Li (2006) extended Rudner's method from decision accuracy to decision consistency.

Alternatively, a simulation-based approach under the IRT framework was proposed by Hambleton and Han (in Bourque, et. al., 2004). Compared to the above analytic approaches, the simulation approach has the merits of being simple to compute, implement and interpret, especially nowadays that there are various IRT generation software packages available and easy to access (e.g, Han & Hambleton, 2007). In addition, the simulation approach is flexible for tests involving complex scoring, scaling, weighting, and equating procedures (e.g., composite scale score). Because there are too many different combinations of subtest scores, it is likely that one raw score will convert to many different composite scale scores. As a result, the analytic expression of the observed score distribution is difficult to identify and compute. Kolen and Harris (2000) pointed out that when the scale transformation is a function of multiple variables,

a simulation approach is preferred (for more discussions on complex assessment, see Brennan & Wan, 2004).

2.3.3. Review of Methods in the Literature

Below is a description for each of six major DC/DA methods: the Huynh extended procedure, the Subkoviak extended procedure, the Livingston and Lewis procedure, the Rudner procedure, the Lee procedure, and the Hambleton and Han procedure. The first three are CTT-based methods, while the last three are IRT-based methods.

2.3.3.1. Huynh and Extended Procedures

Huynh's method assumes a beta distribution with parameters α and β for the true scores, and a bivariate beta-binomial distribution for the observed scores (Keats & Lord, 1962). Let *x* and *y* be the test scores obtained from two parallel forms X and Y. Under the assumption of local independence, x and y follow a bivariate beta-binomial distribution with joint probability density (Huynh, 1976)

$$f(x,y) = \frac{\binom{n}{x}\binom{n}{y}}{B(\alpha,\beta)} B(\alpha + x + y, 2n + \beta - x - y)$$

where B is the beta function with parameter α and β, and *n* is the total number of items. Suppose *C* is the cut score dividing examinees into binary categories, the classification consistency index *P* can be calculated as follows

$$\begin{aligned} P &= P(x \le C-1, y \le C-1) + P(x \ge C, y \ge C) \\ &= \sum_{x=0}^{C-1}\sum_{y=0}^{C-1} P(x,y) + \sum_{x=C}^{n}\sum_{y=C}^{n} P(x,y) \end{aligned}$$

Hanson and Brennan (1990) expanded the model by applying a four-parameter beta

distribution (Lord, 1965) for true score distributions. The four-parameter beta distribution is a generalization of the two-parameter beta distribution that, in addition to the parameters α and β, has two more parameters for the lower (a) and upper (b) limits of the distribution. The true score T follows the distribution with density

$$g(T/\alpha,\beta,a,b)=\frac{1}{B(\alpha+1,\beta+1)}\frac{(T-a)^{\alpha}(b-T)^{\beta}}{(b-a)^{\alpha+\beta+1}}$$

The authors found that the generalized beta-binomial model provided a better fit to the observed score distributions. This is expected of course because additional parameters are available to find the best fitting distribution.

The approach based on the beta-binomial model was mathematically elegant. It was found that the method had comparatively small standard errors (Subkoviak, 1978) and most accurate estimates of P for unimodel distributions (Berk, 1980). Besides, the violation of the equal item difficulty assumption seemed to have negligible effect on the estimates (Subkoviak, 1978). Nevertheless, the method remained one of the most conceptually and computationally complex approaches (Berk, 1980).

To overcome the computational complexity, a simple normal approximation was suggested by Peng and Subkoviak (1980). They found the approximation provided relatively accurate P and *kappa* estimates, and justified the approach with literature which showed that the beta-binomial family could be approximated by the normal family. For low stakes assessment, it has been felt that the Peng and Subkoviak procedure is more than sufficient, and provides a solution that essentially everyone in

psychometrics can apply since it is a simple table look-up.

2.3.3.2. Subkoviak and Extended Procedures

Subkoviak's (1976) method similarly imposed a binomial model on the observed score distributions. However, instead of making distributional assumptions for the true scores, Subkoviak estimated the consistency index for each examinee at a time and then averaged over the examinees. Specifically, Subkoviak estimated each examinee's proportion-correct true score by applying a linear regression approximation using his/her observed proportion-correct scores and the KR-20 coefficient. The conditional observed score distribution was constructed afterwards based on the estimated true score and the binomial model. The consistency index was calculated for each examinee, and then averaged over the sample of examinees. Mathematically, the consistency index for person *i*, defined as $P_c^{(i)}$, is expressed as

$$P_c^{(i)} = P(x_i \geq C)^2 + [1 - P(x_i \geq C)]^2$$

where

$$P(x_i \geq C) = \sum_{x_i=C}^{n} \binom{n}{x_i} \hat{\pi}_i^{x_i} (1-\hat{\pi}_i)^{n-x_i}$$

where $\hat{\pi}_i$ is the estimated proportion-correct true score for examinee *i*, x_i is the examinee's observed score, C is the cut score, and n is the number of items. $P(x_i \geq C)$ is the probability for the examinee in getting a score equal to or higher than the cut score C. The ultimate classification consistency index P was the averaged $P_c^{(i)}$ across the examinees. This method however is highly problematic with short tests

because it leads to poor estimates of domain scores, and when these estimates are zero or 1, the estimates of DC can be far too high.

Lee and his colleagues (Lee, 2007; Lee, Brennan, & Wan, 2009) proposed a compound multinomial (CM) model for a test containing mixture of dichotomous and polytomous items. The compound multinomial procedure can be viewed as a generalized version of Subkoviak's procedure in the sense that it reduces to Subkoviak's procedure when all items are dichotomously scored. A bias-correction procedure (Brennan & Lee, 2006) was applied to make the distribution of observed scores approachable to, having the same amount of variance as, the distribution of true scores.

Brennan and Wan (2004) extended Subkoviak's procedure by developing a bootstrap procedure. Their method is conceptually related to Subkoviak's procedure in terms that it doesn't make distributional assumptions about the true abilities either. By contrast, it generated a large number of replications (called bootstrap samples), and calculated the proportion of consistent decisions for each examinee, and then averaged over examinees. The bootstrap procedure is claimed to be simpler and more flexible for complex assessments when the distribution of observed scores is not easy to estimate. Wan, Brennan and Lee (2007) found that the compound multinomial procedure and bootstrap procedure provided very similar estimates and deemed both as the extension of Subkoviak's procedure.

2.3.3.3. Livingston and Lewis Procedure

Livingston and Lewis (1995) introduced a concept called "effective test length", so that the methods based on the binomial model could be applied to tests which have items polytomously scored or not equally weighted, e.g., the tests containing a mixture of polytomous and dichotomous items, or tests using composite scores. The term "effective test length", denoted as *n*, refers to the number of discrete, dichotomously-scored, locally independent items necessary to produce total scores having the same precision (i.e., reliability) as the scores being actually used to classify examinees. The formula to solve *n* suggested by the authors is

$$n = \frac{(\mu_x - X_{\min})(X_{\max} - \mu_x) - r\sigma_x^2}{\sigma_x^2(1-r)}$$

where $X_{\min}$ is the lowest possible score, $X_{\max}$ is the largest possible score, μ_x is the mean score, σ_x^2 is the test score variance, and r is the classical reliability estimate of the test. It can be displayed from the formula that four kinds of information are required as input: (1) the observed test score distribution, (2) the reliability coefficient of the test, (3) the possible maximum and minimum test scores, and (4) the cut scores.

Using the effective test length, the observed test score *X* ranging from $X_{\min}$ to $X_{\max}$ can be transformed to a new scale X' ranging from 0 to *n* based on $X' = n\frac{X - X_{\min}}{X_{\max} - X_{\min}}$. As Hanson and Brennan (1990) suggested, Livingston and Lewis (1995) estimate the true score distributions by fitting a 4-parameter beta model, and estimate the conditional observed score distributions by fitting a beta-binomial model, based on the estimated effective test length *n.* The Hanson and Brennan procedure, and

the Livingston and Lewis procedure, both based on the beta-binomial model, can be implemented in the software program called BB_CLASS (Brennan, 2004).

2.3.3.4. Rudner Procedure

Rudner (2001, 2005) proposed a procedure for computing classification accuracy indices for both dichotomous and polytomous items in the framework of IRT. In Rudner's approach, the tests are scored on a latent ability scale. θ and $\hat{\theta}$ are denoted as true score and observed score, and θ_c is the cut score. It is assumed that for any true score θ, its corresponding observed score $\hat{\theta}$ follows a normal distribution with mean of θ and standard deviation of SE($\hat{\theta}$). SE($\hat{\theta}$) is the standard error of estimation on θ level. Under the IRT framework,

$$\text{SE}(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$$

$$\text{and} \quad I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

$$\text{and} \quad I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)[1-P_i(\theta)]}$$

where $I(\theta)$ is the test information function, and $I_i(\theta)$ is the item information function. $P_i(\theta)$ is the item response function, and $P_i'(\theta)$ is the derivative of $P_i(\theta)$ with respect to θ (Hambleton, Swaminathan, & Rogers, 1991).

Assuming that $\hat{\theta}$ follows the normal distribution ~N(θ, SE($\hat{\theta}$)), the probability of having $\hat{\theta}$ above θ_c given true score θ, $P(\hat{\theta} > \theta_c \mid \theta)$, is essentially the area under

the normal curve and to the right of $z = \frac{\theta_c - \theta}{SE(\hat{\theta})}$. Taking the binary classifications for example, the classification accuracy index P_A can be expressed as

$$P_A = \int_{\theta=\theta_c}^{\infty} P(\hat{\theta} > \theta_c \mid \theta) P(\theta) d\theta + \int_{\theta=-\infty}^{\theta_c} P(\hat{\theta} < \theta_c \mid \theta) P(\theta) d\theta$$

If one assumes a normal distribution for the true score distribution, $P(\theta)$ is the height of the normal curve at θ. The index P_A can be easily extended to the false positive and false negative error rates.

Rudner focused the attention on DA indices, but DC is a topic of importance too. Li (2006) adapted this approach and extended it to decision consistency. Given the definition that decision consistency is the agreement of classifications across repeated independent administrations, the probability of having $\hat{\theta}$ above θ_c given θ in both administrations is simply the product of probabilities in each administration, that is $P(\hat{\theta} > \theta_c \mid \theta) * P(\hat{\theta} > \theta_c \mid \theta)$. Similarly, the probability of having $\hat{\theta}$ consistently below θ_c given θ in both administrations is the product of $P(\hat{\theta} < \theta_c \mid \theta)$ and $P(\hat{\theta} < \theta_c \mid \theta)$. Still taking binary classifications as an example, the overall decision consistency index P_c is

$$P_c = \int_{\theta=-\infty}^{\infty} P(\hat{\theta} > \theta_c \mid \theta)^2 P(\theta) d\theta + \int_{\theta=-\infty}^{\infty} P(\hat{\theta} < \theta_c \mid \theta)^2 P(\theta) d\theta$$

The logic of DC/DA indices for binary classifications can be easily extended to multiple classifications. The overall procedure follows the logic of Subkoviak's work.

2.3.3.5. Lee Procedure

Different from Rudner's method which is used for tests scored on the θ scale, Lee's (2010) developed a procedure for tests scored by summing up item scores but using IRT as the psychometric foundation. Therefore, it assumes appropriate IRT models are chosen and item parameters are well calibrated.

Lee developed a procedure to estimate the observed summed-score distribution conditional on true ability θ, denoted $P(X \mid \theta)$, and then calculated the consistency index based on the observed summed-score distribution integrated across all examinees. Provided with IRT models and calibrated item parameter estimates, the probability of a vector of item responses given θ can be expressed as

$$P(U_1, U_2, ..., U_n \mid \theta) = \prod_{i=1}^{n} P(U_i \mid \theta)$$

and the probability of a summed score X given θ can be calculated by

$$P(X \mid \theta) = \sum_{U_1,...,U_n : \sum w_i U_i = X} P(U_1, U_2, ..., U_n \mid \theta)$$

where n is the number of items, U_i is the response to item *i*, and w_i is the weight of item *i*. IRT models are used to calculate $P(U_i \mid \theta)$. And the compound multinomial (CM) model was adopted in calculating $P(X \mid \theta)$, where each item was viewed as an item set in this situation. A recursive algorithm was used to compute the compound multinomial model. Kolen and Brennan (2004, pp.219) illustrated the algorithm with examples.

Taking multiple classifications for example, the decision consistency P_c can be formulated as

$$P_c = \int_{-\infty}^{\infty} \sum_{h=1}^{H} [p_\theta(h)]^2 f(\theta) d\theta$$

where H is the number of classification categories, $p_\theta(h)$ is the probability of observed score falling into the h^{th} category conditional on θ, and f(θ) is the density of true score distribution.

Concerning the integration over true scores, two approaches were used. The first one used the estimated quadrature points and weights provided in the IRT calibration output. It was called the D-method by the author since a distributional assumption for true abilities was made. The second approach calculated classification indices for each examinee at a time and averaged over the population. It was called the P-method. The author showed that both approaches produced very similar results.

2.3.3.6. Hambleton and Han Procedure

The above methods were developed based on analytical approaches. Some of the modelings are very complicated and the computation are not easy or straightforward. Along with wide spread applications of IRT and the availability of a number of IRT software programs for calibration and generation, Hambleton and Han (in Bourque et. al., 2004) proposed a convenient and straightforward method based on Monte-Carlo simulation techniques. This simulation-based method was initially suggested for dichotomous data, but can easily be extended to polytomous data. It makes no assumptions about the score distributions, except that the data fit the IRT models and satisfy IRT model assumptions, e.g., dimensionality and local independence. These

assumptions are fair to make and convenient to check because they are the prerequisites and always need to be checked prior to any IRT applications.

According to the authors, the inputs are (1) item and ability parameter estimates, which can be calibrated given the response data and chosen IRT models, and (2) the cut scores. A classical test reliability coefficient estimate can be provided to correct the true score distribution using the Kelley regressed estimates (Kelley, 1947). The correction has been shown having minimal impacts on the DC/DA results by Li (2006) and was skipped in this study. The method can be described in a three-step procedure:

(1) Generate test response data.

Provided with item parameter and ability estimates, and an appropriately chosen IRT model, two sets of response data for parallel form X and Y are generated. Calculate the test scores for form X and Y.

(2) Transforming cut scores and classifying examinees.

Use the test characteristic curve (TCC) to transform the cut scores to the test score metric if they are provided on the theta metric. The TCC can be constructed using the available item parameter estimates. Classify examinees into performance categories based on their test scores on form X and Y and the cut scores on the test score metric. Classify examinees into "true" performance categories based on their ability estimates and the cut scores on the theta metric.

(3) Calculate the classification indices.

Calculate the DC indices based on the classifications of examinees using parallel

form scores on X and Y. Calculate the DA indices based on the classifications using ability estimates and the test scores of one of the parallel forms. Alternatively, the average of the two possible values of DA can be used as the single estimate of DA.

The Hambleton and Han method is easy to understand, to calculate, and to interpret. It avoids complicated models and daunting computations by generating test scores for parallel forms from fitted IRT models. It simply calculates DC indices based on the degree of classification agreement using the scores of parallel forms, and using the ability scores for DA indices. One disadvantage is that the indices' values may vary a bit from one calculation to another since it is based on simulation, and variation of DC and DA statistics would be expected. It is suggested that the simulation be performed multiple times and choose the mean of the values across the replications. Alternatively, it suggests simulating large samples so that precise estimates of DC and DA can be obtained from a single simulation. Large samples do not have implications for any aspects of the study except the stability of the DC and DA estimates.

2.3.4. Summary and Conclusion

Two tables were provided summarizing the major methods for estimating DC and DA indices using a single administration, Table 2.1 for CTT-based methods and Table 2.2 for IRT-based methods. The methods were described in terms of their sources, features, assumptions, and whether they are applicable to polytomous data or not.

Berk (1980) argued that whenever parallel forms were available, the two-administration approach was recommended over single-administration because it

was unbiased and more accurate. Nevertheless, the single-administration estimate is popular in practice due to its convenience and availability. However, few studies are available and these methods are not well studied. Among the few simulation studies in the literature, Wan, Brennan, and Lee (2007) conducted one study examining the performances of four CTT-based methods in various conditions using both simulated and real data. They found that generally Livingston & Lewis (1995), Peng & Subkoviak (1980) methods outperformed Brennan & Wan (2004) and Lee (2005) methods. However, LL and PS methods were suspected to be more sensitive to reliability estimates, and to score distributions. Besides, the LL method tended to substantially underestimate kappa for certain cut scores when the correlation between constructs was not 1.0. The reason was however not clear.

Lee, Hanson, and Brennan (2002) conducted another comparative study comparing the performance of methods assuming three different models: the two-parameter beta binomial model (2PB), four-parameter beta binomial model (4PB) and the three-parameter logistic IRT model (3PL). The study used real data and the examinees were scored using the number-correct method. The authors found that the 3PL model fitted better to the data than 4PB model, and in turn better than the 2PB model. The *P* estimate did not differ greatly across the models while *kappa* varied more substantially, and the 3PL model yielded the highest values of the indices. In addition, the 4PB yielded severely skewed true score distributions. However, it was argued that since "true" values and true score distributions were unknown, no conclusion could be drawn

concerning which estimate was more accurate. The authors called for a comprehensive simulation study. It is noteworthy that none of the above studies in literature examined the *P* estimate for decision accuracy in a simulation study.

Li (2006) evaluated three IRT-based methods (Rudner, adapted Rudner, and Hambleton and Han methods) and compared them with the Livingston and Lewis method. Their robustness to various test lengths and true score distributions was examined in a series of simulation studies. She found the three IRT-based methods performed satisfactorily most of the time, and slightly better than LL method. She also found that the HH and LL methods were more sensitive to short tests. The RD method was comparatively robust to skewed ability distributions. As for DC/DA estimates, the DA has more accurate estimates than DC, and in turn than *kappa*. Rudner had the highest values of the indices, L&L the lowest, and H&H was in between. Of course the issue here is not to generate high values of DC or DA but rather the goal is to produce accurate estimates of DC and DA.

Decision consistency and accuracy indices have been routinely reported in many testing programs today. Some CTT-based methods, especially the Livingston and Lewis (LL) method, have been widely used in practice in reporting the DC and DA indices. However, the methods are not widely studied. One obvious problem is that the LL method is sensitive to the choice of test reliability estimate. Given that several reliability estimates are available in the literature, no study has been shown to discuss which reliability estimate is a better choice or what the practical differences are in the DC and

DA estimates as a function of the choice of reliability estimate. Would a difference between .85 and .90 in the reliability estimate make a difference in the DC and DA estimates, for example? Recall that .85 might arise if the KR-20 value is used, and .90 might arise if parallel-form reliability estimate is used.

New IRT-based methods were developed recently and these methods deserve further study too. It would be of great interest for the researchers and practitioners to know how accurately these methods perform and how robust they are to non-standard test conditions. It is disappointing to find that few studies existing in the literature addressed these questions. Given the deficit that the "true" values of indices and the "true" score distribution are unknown in the real data, a comprehensive study with both simulated and real data is therefore greatly desired. It is hoped that a study like this one will help better understand these methods and their variations in special test conditions and how to choose an appropriate method in practice.

Table 2.1 CTT-based Methods of Estimating DC/DA Index

Source	Feature	Assumption	Poly Data
Huynh (1976)	2-parameter beta-binomial model	Beta-binomial model assumed for observed score distribution; beta distribution assumed for true score distribution	
Hanson & Brennan (1990)	4-parameter beta-binomial model		
Livingston & Lewis (1995)	4-parameter beta-binomial model, effective test length.		√
Subkoviak (1976)	Binomial model, consistency index estimated for each person and averaged across persons	Binomial assumed for observed score distribution, no assumption for true score distribution	
Brennan & Wan (2004)	Bootstrap procedure	Compound binomial/multinomial model assumed for observed score distribution, no assumption for true score distribution	√
Lee (2007) Lee, Brennan & Wan (2009), Brennan & Lee (2006)	Compound multinomial model, bias correction for true score distribution		√
Peng & Subkoviak (1980)	Normal approximation of beta binomial distribution	Bivariate normal model assumed for observed score distribution	√
Woodruff & Sawyer (1989)	Split-half approach, Spearman-Brown formula applied		√
Breyer & Lewis (1994)	Split-half approach, tetrachoric correlation used, Spearman-Brown formula applied		√

Table 2.2 IRT-based Methods of Estimating DC/DA Index

Source	Feature	Assumption	Poly Data
Huynh (1990)	Rasch model	IRT models and compound binomial/multinom-ial model	
Wang, Kolen, and Harris (2000)	Polytomous IRT models		√
Lee (2010)	Mixture IRT models		√
Rudner (2001, 2005), Li (2006)	Test scored on theta scale	IRT models	√
Hambleton & Han (in Bourque, el. at., 2004)	Simulation-based approach	IRT models	√

		Administration 1	
		Failing	Passing
Administration 2	Failing	p_{00}	p_{01}
	Passing	p_{10}	p_{11}

Figure 2.1 Agreement Index for Decision Consistency

CHAPTER 3

SIMULATION STUDIES

3.1. Method

3.1.1. Selected DC/DA Methods

Four variations of three methods were selected for investigation. The three methods are the Livingston and Lewis method, Lee method, and the Hambleton and Han method, denoted as LL, LEE and HH, respectively. The LL method is CTT-based and is currently the most popular method. Despite its popularity, there are not many studies available on the LL method. Some literature (Wan, Brennan, & Lee, 2007) suspected its sensitivity to factors such as the reliability, skewed distribution, location of cut scores, etc. The LEE and HH methods are recently developed IRT-based methods, which deserve further investigation too since they are relatively new and unstudied.

The software BB-CLASS (Brennan, 2004) and IRT-CLASS (Lee & Kolen, 2008) were used to implement the LL and LEE methods. The HH method was programmed by the author using R.

In addition, a variation of the LL method, using a stratified version of coefficient alpha as the reliability estimate in the input rather than the standard coefficient alpha, denoted as LL_{Strat}, were investigated too. The reliability estimate is a major input in LL method, and the results of LL method are sensitive to the choice of reliability estimate (a higher reliability estimate results in a higher DC/DA estimate while other inputs being equal). It is of significance to explore which reliability estimate is a better choice

provided that several choices exist in the literature. The most popular standard Cronbach's alpha assumes that all the items are parallel and independent. In real settings, it is, however, natural to expect that if the items were divided into categories (based on item type, item content, etc.), a unique variance is associated with the categories. The stratified coefficient alpha was proposed treating the items in one category as a separate subtest (Rajaratnam, Cronbach, & Gleser, 1965). It is given by

$$\alpha_{strat} = 1 - \frac{\sum_{j=1}^{k} \sigma_{xj}^2 (1 - \alpha_j)}{\sigma_x^2}$$

where σ_{xj}^2 is the variance of scores of the j^{th} category, α_j is the standard Cronbach's alpha in the j^{th} category, and σ_x^2 is the total test score variance. In this study, the stratified version of alpha was used as an alternative in the LL method where the reliability estimated is needed. For the sake of convenience in implementation, it was calculated by dividing the items based on their item types (dichotomous vs. polytomous items).

3.1.2. Data

The three-parameter logistic (3PL) IRT model (Birnbaum, 1968) and the two-parameter graded response model (GRM) (Samejima, 1969) were used to generate the standard unidimensional data for the dichotomous and polytomous items, respectively. The 3PL model is the same as presented in the previous chapter. The two-parameter GRM is given by

$$P_{ix}(\theta) = P^{*}_{ix}(\theta) - P^{*}_{i(x+1)}(\theta)$$

$$\text{where} \quad P^{*}_{ix}(\theta) = \frac{e^{Da_i(\theta - b_{ix})}}{1 + e^{Da_i(\theta - b_{ix})}}$$

$P_{ix}(\theta)$ is the probability of examinee with ability θ getting a score category x in the polytomous item i. $P^{*}_{ix}(\theta)$ is the probability of examinee with ability θ getting a score category x or above, which is essentially the two-parameter logistic (2PL) model. a_i is the item discriminating parameter, b_{ix} is the threshold parameter, or category boundary, for the score category x. And D is the scaling consistent ($D = 1.7$).

The 3PL and GRM are two popular IRT models which are widely used and often have good fit with standardized tests (e.g., the MCAS tests). The two models were therefore chosen in generating the data to mimic the real situation. It was of interest to study how the selected methods would perform in various conditions when the data fit the 3PL/GRM models.

To study the impacts of local item dependence on DC/DA estimates, the Testlet Response Theory (TRT) models were used to generate the data with local item dependency. The concept of testlet refers to a group of items which share common stimuli, are content- or format-dependent, and are interdependent with each other. The testlet effect is often viewed as a secondary dimension besides for the dominant dimension of true ability, and is a common threat to the fundamental assumptions of standard IRT models: unidimensionality and local item independence.

The TRT models (Wainer, Bradlow, & Du, 2000) are a modification of the standard

unidimensional IRT models. An example of the 3PL TRT model is given below:

$$p(y_{ji}=1)=c_i+(1-c_i)\frac{\exp[a_i(\theta_j-b_i-\gamma_{jd(i)})]}{1+\exp[a_i(\theta_j-b_i-\gamma_{jd(i)})]}$$

where $p(y_{ji}=1)$ is the probability of examinee *j* answering item *i* correctly. The parameters a_i, b_i, and c_i are identical as in the standard 3PL model. $\gamma_{jd(i)}$ is the additional interaction term introduced to the standard IRT models. It reflects an interaction between person *j* and the testlet *d(i)* which contains item *i*. $\gamma_{jd(i)}$ can be interpreted as the examinee's ability in the secondary dimension associated with the testlet *d(i)* which is unrelated with the dominant dimension of the underlying latent ability (e.g., the candidate's background knowledge with the content of the reading passage in a reading comprehension test). (Note that the TRT model is essentially a special case of the full-information item bifactor model. The bifactor model is also called the general testlet model, where the testlet effect is treated as a group factor. The difference is that the TRT model imposes a constraint on the *a*-parameter by assuming the discriminating powers associated with latent trait (θ_j) and with testlet effect ($\gamma_{jd(i)}$) are equal, whereas the bifactor model does not impose such a restriction and the discriminating powers can be different.)

Although $\gamma_{jd(i)}$ is the testlet effect parameter, its variance $\sigma^2_{d(i)}$ rather than itself, indicates the degree of local item dependency within testlet *d(i)*. $\sigma^2_{d(i)}$ can be varied across different testlets. A larger value of $\sigma^2_{d(i)}$ indicates a higher level of local

dependency, vice versa. A $\sigma^2_{d(i)}$ of zero indicates the items within the testlet are essentially locally independent.

The 3PL and GRM TRT models were used to generate dichotomous and polytomous data, respectively. For the sake of easy implementation, the item type effect was treated as the testlet effect in simulation. There were two testlet effects in the test, γ_{mc} associated with all multiple-choice items, and γ_{fr} associated with all free-response items. Both γ_{mc} and γ_{fr} followed a normal distribution, with mean of zero for the purpose of identification. The latent ability θ, γ_{mc} and γ_{fr} were uncorrelated with each other.

The "true" theta scores were generated from a normal distribution with a mean of zero and SD of one. The "true" item parameters were randomly drawn from the item pool of a real statewide standardized achievement test (MCAS ELA grade 10 in 2009), which had 84 multiple-choice items (scored 0/1) and 12 free-response items (scored 0-4). The tests of various test lengths were created by randomly drawing specified numbers of items from the pool. The data were created in this way so that they mimicked the real situations and the generalizability of the findings would be enhanced.

Adopted from the real test, three cut scores were used to classify examinees into four proficiency categories: the failing (F), needs improvement (NI), proficient (P), and advanced (A) levels. The three cut scores were provided on the theta scale so that the "true" classifications were able to be calculated, provided that the "true" theta scores were known in the simulation studies. The cut scores were set at -1.75, -0.81, and 0.58,

which classified the candidates with the percentages of 4% (F), 17% (NI), 51% (P), and 28% (A) in each of the proficiency levels, about the same percentages observed in the real test.

It is noteworthy that because the LL and LEE methods score and classify examinees based on their raw scores, the generated data were scored on the raw score scale in the first three simulation studies so that the performance of the selected methods were comparable. To classify examines on raw score scale, the cut scores provided on the theta metric were converted to the raw score metric using the test characteristic curve. One exception was Study 4, where the data were scored on the theta and composite score metrics for applicable methods so that the impacts of using different scoring methods were able to be assessed.

3.1.3. "True" DC/DA Indices

The "true" DC/DA indices were calculated so that the accuracy of DC/DA estimates from selected methods could be evaluated. The diagram in Figure 3.1 displayed the procedures in calculating the "true" indices. Specifically, the following three steps were followed:

(1) The data were scored and classified, and the classification observed from the data was called "actual" classification. For example, if the data were scored on raw score scale, the theta cut scores were converted from theta scale to raw score scale using test characteristic curve, and candidates were classified based on their raw scores. Based on the metric or method chosen for the data, an appropriate DC/DA method was applied

to the data, which gave the DC/DA estimates. The DC/DA estimates are estimating how accurate and consistency the "actual" classification is.

(2) The DA is defined as the degree of agreement between the classification based on the data and the classification based on the candidates' true scores if the true score were known. The true scores are not possibly known in reality, but were known in simulation. In Step 2, the candidates were classified based on their true theta scores, and the classification was referred to as "true" classification. "True" PA was simply the degree of agreement between the "true" classification and the "actual" classification, the latter one was described in Step 1.

(3) The DC is defined as the degree of agreement between the classification based on the data and the classification based on its parallel form. In Step 3, provided that both "true" theta scores and "true" item parameters were known in simulation, a strictly parallel form was able to be generated, labeled as Data 2 in the diagram. This second parallel data were scored and classified following exactly the same procedure as for the original data as described in Step 1.The observed classification from the second data was called "actual" classification 2. "True" PC is the degree of agreement between the "actual" classifications observed from the two parallel data sets. The "true" Kappa was computed accordingly based on the contingency table.

3.1.4. Evaluation Criterion

BIAS was calculated as the criterion to examine how accurate each of the selected methods was in various conditions. BIAS reflects both the systematic error (by the sign

of statistic) and the random error (by the absolute value of the statistic) of the estimates. It is given by

$$BIAS(\hat{P}) = \hat{P} - P$$

where P is the "true" DC/DA index, and $\hat{P}$ is the DC/DA estimate. Usually $\hat{P}$ takes the mean of estimates across a number of replications. In this study, the sample size was eliminated as a factor. By using a large sample size of 50,000 the sampling errors could be minimized without replications, and the estimate was deemed as the mean across a couple of replications. Specifically, $\hat{P}$ was replaced by $\hat{P}_C$, $\hat{P}_A$ and $\hat{K}$ to represent the estimates of PC, PA and Kappa, respectively.

3.2. Study 1: Robustness to Test Length

3.2.1. Purpose of the Study

The test length was considered as a factor which would impose an impact on the DC/DA estimates for two reasons. Firstly, the test length has a direct impact on test reliability. Given items of the same quality, the shorter the test is, the lower the test reliability will be. Remembering that the reliability estimate is a major input in the LL method, and the LL method was suspected to be sensitive to reliability estimates (Wan, Brennan, & Lee, 2007). Secondly, the test length would have an impact on the ability estimates in the IRT framework. The shorter the test is, the larger the errors are in the ability estimates.

Study 1 addressed two questions:

(1) What is the impact of short test length on "true" DC/DA indices?

(2) How robust are the selected methods to short tests?

3.2.2. Conditions

Four test length conditions were considered: 10/1, 20/2, 40/4, 80/8. The numbers before the slash denoted the total number of items, while the numbers after the slash denoted the number of polytomous items. The proportion of polytomous items was fixed in order to eliminate the possible effects of proportion of polytomous items on the DC/DA estimates. The reliability estimates of the various test lengths were around 0.75, 0.85, 0.9 and 0.95, respectively, and these reliabilities are certainly in the range seen in practice though .95 would be judged as rather high in practice but perhaps seen with some Advanced Placement (AP) Tests. The "true" item difficulty parameters drawn from the real test for 4 various test lengths were summarized in Table 3.1. The table showed that they were a bit difficult tests, and this was especially true for short tests. It is worthwhile noting that since the condition of 10 items had only one item in the category of polytomous items, the stratified alpha cannot be computed. For the same reason the LL method using stratified alpha was not applicable to this condition.

3.2.3. Results

3.2.3.1. Reliability Estimates

The two alternative choices of reliability estimate, the Cronbach's coefficient alpha and stratified alpha, were calculated for the tests of four different lengths and were summarized in Table 3.2. The stratified alpha produced a slightly higher value of reliability estimate than the Cronbach's alpha. The increase was around 0.02 for short

test of 20 items, and was negligible for long test of 80 items.

3.2.3.2. "True" DC/DA

Table 3.3 and Figure 3.2 displayed the "true" PA, PC and Kappa indices in various test lengths in Study 1. The tables and plots showed that when the test had normal ability distribution, a longer test had higher values of "true" PA, PC, and Kappa indices than a shorter test.

3.2.3.3. Bias

Table 3.4 and Figure 3.3 included the biases for selected methods in various test lengths. The results indicated that (1) all methods had reasonably good performance and small bias (absolute value smaller than 0.05) across different test lengths. Although it is obvious for all methods that the bias was decreased as the test got longer. (2) The LL_strat, LEE and HH methods had smaller bias than LL method in most conditions. The LL method seemed to have poorer estimates and was more vulnerable, especially, in short tests. Besides, the results indicated that the LL method consistently under-estimated DC/DA estimates.

3.3. Study 2: Robustness to Local Item Dependency

3.3.1. Purpose of the Study

Both the IRT- and CTT-based methods assume that the items are independent when the true score is controlled for. However, sometimes some items in the test are interrelated with each other due to various reasons, e.g., sharing a common content or format, and the consequence is called local item dependency (LID). It is not unusual in

practice that a standard unidimensional IRT model is applied to the data with LID. In this study, it is of interest to investigate what are the impacts of LID when it comes to DC/DA indices. And does the classically-based procedure function better than the IRT-based procedure? Specifically, it was intended to answer the following questions:

(1) What are the "true" DC/DA indices when tests with LID are calibrated using unidimensional IRT models?

(2) How accurate are the selected methods with tests of various degrees of LID?

3.3.2. Conditions

The 3PL/GRM TRT models, as discussed in the previous chapter, were used to generate the data. Two testlet effects were generated to create the local item dependency, one associated with all multiple-choice items and denoted as γ_{mc}, and the other associated with all free-response items and denoted as γ_{fr}. Two factors were considered in generating the data:

(1) The degree of local item dependence within testlets. This factor was manipulated by setting $\sigma^2_{d(i)}$, the variance of testlet effects across persons, to different values: 0, 0.2, 0.5, and 1, where 0 indicated no local dependency within the testlets, and 1 indicated a high level of local dependency within the testlets.

(2) The number of items associated with each testlet effect. Simulations showed that the more disparate the numbers of items in different testlets are, the more dominant the first factor is; on the contrary, the closer the numbers are, the stronger the second factor is, with other conditions being equal. Tests of two different lengths were

generated: 40 items, consisting of 36 multiple-choice (MC) items and 4 free-response (FR) items, and 36 items, consisting of 28 MC items and 8 FR items. All FR items were scored score 0 to 4. The two test lengths were compared because they had about the same test reliability estimates, 0.924 and 0.925, separately. With items being equal, the first condition had a stronger first factor than the second condition, because the number of items in one testlet (that was, the MC items) was more dominant.

To summarize, 8 conditions, 2 test lengths crossed by 4 degrees of local dependency with testlets, were studied in Study 2.

3.3.3. Results

3.3.3.1. Dimensionality Analysis

The dimensionality of the tests with various degrees of local dependency was analyzed using Principle Component Analysis (PCA). Table 3.5 provided the numbers of eigenvalues for the eight test conditions, while Figure 3.4 and Figure 3.5 displayed the eigenvalue plots. The tables and plots showed that the eight tests exhibited different degrees of dimensionality, from very unidimensional to moderately multidimensional. As $\sigma^2_{d(i)}$ got larger, the first factor became weaker. In addition, the test with 28 MC and 8 FR had a stronger second factor than the test with 36 MC and 4FR. Both of them were as expected. The table also suggests that the ratio of the first to the second factors be a better criterion in judging unidimensionality than the proportion the first eigenvalue accounting for.

3.3.3.2. "True" DC/DA

Table 3.6 to Table 3.8 provided the "true" values and the PA, PC, and Kappa estimate separately. The "true" indices were plotted in Figure 3.6. The plots showed that: (1) the degree of LID had a negative impact on "true" PA, which dropped by about 0.2 as the variance of testlet effects increased from 0 to 1. (2) The degree of LID did not have an impact on "true" PC or Kappa. (3) The "true" PA was larger than "true" PC only when the data was unidimensional, but not in the presence of various degrees of LID. (4) The two different test lengths in this study did not show an obvious impact on "true" DC/DA indices. Although the test of 28MC + 8FR had slightly higher "true" PC /Kappa than test of 36 MC + 4FR, the differences were trivial. It might because the reliability with 28MC + 8FR was slightly higher but other explanations could be possible.

3.3.3.3. Bias

Table 3.9 to Table 3.11 give the bias of estimates and the plots are displayed in Figure 3.7 and Figure 3.8. The biases which had an absolute value larger than 0.05 were highlighted in bold and italics in the tables. Observation of the results suggests that (1) LID had a negative impact on PA estimates. Using a criterion of 0.05, none of the methods produced small bias when the tests had from minor to high levels of local dependency. All methods over-estimated PA when $\sigma^2_{d(i)} > 0$, and the over-estimation was severer as $\sigma^2_{d(i)}$ got bigger. The largest bias reached around 0.24 for test of 28MC + 8FR with $\sigma^2_{d(i)}$ =1. (2) The methods produced satisfactory bias for PC/Kappa in most

conditions, except that the IRT-based methods with $\sigma^2_{d(i)} = 1$ had a bias larger than 0.05. It seems that the IRT-based methods were more vulnerable to a higher level of LID while the LL_strat method performed the best among the four method options. (3) The biases for tests of 28MC+8FR were larger than that of 36MC+4FR, conditioning on $\sigma^2_{d(i)}$. It should be easy to explain as 28MC+8FR had a stronger secondary factor.

3.4. Study 3: Robustness to IRT Model Misfit

3.4.1. Purpose of the Study

Even though the data meet all the requirements underlying the models, the methods were put at a risk of malfunctioning if an incorrect model was chosen to fit the data. The misfit could happen due to the fact that the procedure of checking model fit has been skipped, or a simpler IRT model is chosen for the sake of convenience, cost, or availability of software, etc. The model-data misfit usually cast negative impacts on IRT applications, and there is no excuse to have an exception for the DC/DA estimates. In Study 3, the scenario of IRT model-data misfit was simulated, and the performance of the selected methods in the presence of model misfit was investigated.

3.4.2. Conditions

Two conditions were compared by fitting both the misfitting and correct models to the data. The consequences of misfitting model on DC/DA estimates were then checked. The two sets of IRT models fitted were (1) the 1-parameter logistic (1PL) model for dichotomous items and partial credit model (PCM) for the polytomous items, (2) 3PL model for the dichotomous items and the GRM for polytomous items.

The 3PL/GRM models were the correct models since they were used in data generation. Then the 1PL/PCM models were used to fit to the data, which mimicked the scenario commonly observed in some testing programs. Only the test condition of 40 items with a normal ability distribution was looked at in this study for the purpose of convenience in interpreting the results. All the selected methods were evaluated in terms of the consequences of IRT model misfit on DC/DA estimates. It is noteworthy that although the LL-based methods, LL and LL_{strat}, were not developed in the IRT framework, the raw cut scores in their input were converted from theta cut scores using the test characteristic curve. Therefore it is still of interest to check whether there would be any possible impact on their DC/DA estimates.

3.4.3. Results

3.4.3.1. "True" DC/DA

The "true" indices of using two sets of IRT models were displayed in Table 3.12 and plotted in Figure 3.9. It showed that compared with 3PL/GRM, 1PC/PCM had a slightly lower "true" PA, about 0.02 lower, but a slightly higher "true" PC, around 0.06 higher. It was as expected that "true" PA decreased when fitted with 1PL/PCM because the validity of method was challenged with a wrong model. However, it was not clear why fitting 1PC/PCM would result in a higher value of "true" PC, and it did show some practical difference.

3.4.3.2. Bias

Table 3.13 displayed the bias of PA/PC/Kappa, and Figure 3.10 showed the plots

accordingly. The results showed that (1) All methods overestimated PA in the condition of 1PL/PCM. The biases were, however, under 0.1. (2) All methods well estimated PC and Kappa indices. It seemed that the misfit between 1PL/PCM models and 3PL/GRM data had some impacts on the accuracy of PA estimates but minimal impacts on the accuracy of PC and Kappa estimates.

3.5. Study 4: Robustness to Scoring Metric

3.5.1. Purpose of the Study

A test can be scored and reported in various ways. For example, the examinees can be scored using raw scores or scale scores. In addition, the abilities can be estimated using theta scores in the IRT framework. Raw scores usually refer to the number of items correct, or the numbers of points earned, without any transformation, e.g., if an examinee answers 50 out of 100 items correctly, with one score point for each item, his/her raw score is 50. Scale scores are transformed from raw scores based on some relationships for the purpose of reporting and interpretation convenience. The relationship can be linear or nonlinear, and both are common in practice. Theta scores refer to the latent trait scores estimated with IRT models which are usually placed on a scale with a mean of 0, and SD of 1. The typical values of thetas vary from -3 to 3. The transformation between theta scores and raw scores can be obtained by using the test characteristic curve function (TCC). More complex scoring includes the composite score, which is a weighted sum of score on two or more subtests. The weights may be equal or unequal. For example, a summed score giving equal weights to multiple-choice

items and constructed-response items was employed by the bar examinations (Wan, Brennan, and Lee, 2007).

When examinees are scored on a particular scale, the cut scores should be transformed and the DC/DA indices should be estimated accordingly. Some DC/DA methods were developed for certain scales while others were developed for different scales. Among the selected methods, the LL method can be used for raw score scale but not for the theta scale. Although the authors of the LL method claimed that the method can be applied to scores on any scale as long as an appropriate reliability coefficient can be provided, the calculation of reliability estimate for theta scores require further efforts and was not studied much. The LEE method was developed for tests scored using raw or scale scores under the IRT framework. The HH method used a simulation approach without making distributional assumptions for the reported scores. Therefore it is flexible and should be applicable for scores reported on any metrics.

A method performing well in one situation may not perform equally well in another in which the method was not originally developed. Study 4 was designed to check how the selected methods would perform for tests scored on different metrics. Specifically, the purposes of this study were to address two questions:

(1) Does it matter that the examinees are scored on raw, theta, or composite score with respect to "true" DC/DA indices?

(2) In addition to the previous studies in which the examinees were scored on raw scores, how accurate were the selected methods for the theta and composite scores?

3.5.2. Conditions

The data generated in Study 1 were scored on two metrics: (1) theta score and (2) composite score. The "true" DC/DA indices and their estimates on the theta and composite scores were calculated and compared with that on the test (raw) score metric.

Different methods were selected in investigation for different metrics. The HH method was the only method for the theta scores, since the other three methods were not applicable. The LL, LL_{strat}, and HH methods were used for the composite scores. The LEE method was not applicable of composite score. Since it uses a recursive algorithm (see Kolen & Brennan, 2004, pp. 219 for examples) to calculate the probability of each possible value of reported scores, there would be too many combinations of subtest scores for each composite score and the computations become very demanding. Therefore the LEE method can only be implemented with a simple scale transformation with the current software, e.g., the one-to-one raw-to-scale conversion.

To score and classify examinees on the theta score, the procedure was straightforward. The data were calibrated using the 3PL/GRM models, and the cut scores provided on the theta scale were applied to the theta scores directly.

To score and classify examinees on the composite score, the procedure was a bit more complex. The formula for composite scores was given by

$$\text{Composite score} = W_{MC} * X_{MC} + W_{FR} * X_{FR}$$

where W_{MC} and W_{FR} are the weights applied for each score point for the MC and FR

items, respectively. X_{MC} and X_{FR} are summed raw scores for the MC and FR items, respectively. Adopting from the real data provided by the Advanced Placement tests, the W_{MC} and W_{FR} were defined so that the weighted sum scores for MC and CR items contribute 60% and 40%, respectively, to the weighted total sum score. The weights for different tests were summarized in Table 3.14. To calculate the cut scores on the composite score scale, the cut scores provided on the theta score were converted to the raw scores for MC items using the test characteristic curve consisting of all MC items, and for FR items using the test characteristic curve consisting of all FR items, separately. Then the weighted cut scores on raw score scale for the MC and FR items were summed up to obtain the final cut scores on composite score scale.

3.5.3. Results

3.5.3.1. "True" DC/DA

The "true" DC/DA indices on three different scoring metrics were summarized in Table 3.15 to Table 3.17, and were plotted in Figure 3.11. The plots showed that (1) the classification based on raw score and on composite score had close "true" PA, PC, and Kappa indices in tests of different test lengths. (2) The classification based on theta score had higher PA, PC and Kappa estimates than that on raw score when the test was short, however, the difference disappeared when the test got longer.

3.5.3.2. Bias

The biases of estimates on three different scoring metrics were summarized in Table 3.18 to Table 3.20. Figure 3.12 displayed the bias on composite score for the LL,

LL_{strat}, and HH methods. The plots showed that when the composite score was used for classification, the three methods all had small bias for PA estimate, however, the LL method appeared to have large bias for PC and Kappa estimates in short tests. In addition, the HH method had the most accurate DC/DA estimates compared to the other two methods and in all test length conditions.

Figure 3.13 plotted the bias on the theta score scale for the HH method. The plots showed that the HH method had small bias and the DC/DA indices were well estimated when the theta score was used for classification. In addition, the biases became smaller when the test got longer.

3.6. Summary and Conclusion

The LL, LL_{strat}, LEE and HH methods were investigated in a variety of conditions in Chapter 3 using a series of simulation studies. Four studies were designed to evaluate the performance of the selected methods in different conditions including various test lengths, local item dependency, model misfit, and different scoring metrics. In addition, the impacts of these conditions on "true" DC/DA indices were checked also.

The findings showed that a longer test had higher values of "true" PA, PC, and Kappa indices. All methods had reasonably small biases across different test lengths, however, the LL method had larger biases when the test was short.

Both LID and IRT model misfit had noticeably decreased the "true" PA index, and caused PA was over-estimated by all selected methods. The worst case for PA was being over-estimated by 0.25 when the test had a high level of LID. On the contrary,

either LID or model misfit exhibited an obvious impact on “true” PC or Kappa indices. In addition, the study found that the IRT-based methods were less robust in PC and Kappa estimates when the test had a high level of LID.’

The scoring metric did not have an apparent impact on DC/DA indices. Although the “true” indices appeared higher for theta score than for raw and composite scores, the differences diminished when the test got longer. Similar with raw score, the LL method had larger bias when the test was short for composite score. The HH method was found performing consistently well across different scoring metrics.

In addition, it was found that the LL method had kept under-estimating PC/Kappa by various degrees in all conditions, and the LL_{strat} method noticeably improved the estimates especially when test was short or had LID.

Table 3.1 Descriptive Statistics of “True” Item Difficulty Parameters

Test Length	Mean	SD	Skewness
10	0.269	0.643	1.074
20	0.308	0.820	0.517
40	0.113	0.808	0.046
80	0.145	0.809	0.387

Table 3.2 Reliability Estimates of Different Test Lengths in Study 1

Test Length	Cronbach’s Alpha	Stratified Alpha
10	0.73	
20	0.85	0.87
40	0.92	0.93
80	0.96	0.96

Table 3.3 “True” and Estimated DC/DA Indices in Study 1

Index	Test Length	"True” Index	LL	LL_Strat	LEE	HH
PA	10	0.6805	0.6611		0.6599	0.6602
	20	0.7543	0.7393	0.7646	0.7438	0.7430
	40	0.8273	0.8151	0.8275	0.8246	0.8267
	80	0.8778	0.8722	0.8825	0.8757	0.8764
PC	10	0.6246	0.5852		0.6142	0.6148
	20	0.6802	0.6565	0.6857	0.6695	0.6691
	40	0.7638	0.7458	0.7614	0.7603	0.7619
	80	0.8266	0.8197	0.8338	0.8265	0.8291
Kappa	10	0.4147	0.3566		0.3969	0.3986
	20	0.5118	0.4734	0.5184	0.4947	0.4941
	40	0.6302	0.6062	0.6305	0.6290	0.6313
	80	0.7266	0.7153	0.7377	0.7258	0.7298

Table 3.4 Bias of DC/DA Estimates in Study 1

Index	Test Length	LL	LL_Strat	LEE	HH
PA	10	-0.0194		-0.0206	-0.0203
	20	-0.0149	0.0103	-0.0105	-0.0113
	40	-0.0122	0.0002	-0.0027	-0.0006
	80	-0.0055	0.0047	-0.0021	-0.0014
PC	10	-0.0395		-0.0104	-0.0098
	20	-0.0238	0.0054	-0.0107	-0.0112
	40	-0.0180	-0.0024	-0.0035	-0.0019
	80	-0.0069	0.0073	0.0000	0.0025
Kappa	10	***-0.0581***		-0.0178	-0.0161
	20	-0.0384	0.0066	-0.0171	-0.0178
	40	-0.0239	0.0003	-0.0011	0.0011
	80	-0.0113	0.0110	-0.0009	0.0031

Table 3.5 Eigenvalues of Tests Conditions in Study 2

Test Length	Testlet Effect Variance	% of 1st Eigenvalue	Ratio of 1st to 2nd Eigenvalues
36MC + 4FR	0	38.1	17.2
	0.2	39.3	14.7
	0.5	41.3	9.5
	1	44.3	7.4
28MC + 8FR	0	41.8	17.1
	0.2	41.6	9.4
	0.5	41.7	5.3
	1	42.7	3.7

Table 3.6 “True” and Estimated PA in Study 2

Test Length	Testlet Effect Variance	"True" PA	LL	LL_Strat	LEE	HH
36MC + 4FR	0	0.8246	0.8163	0.8286	0.8241	0.8252
	0.2	0.7520	0.8187	0.8341	0.8201	0.8195
	0.5	0.6769	0.8072	0.8318	0.7949	0.7947
	1	0.6141	0.8127	0.8415	0.7748	0.7753
28MC + 8FR	0	0.8489	0.8404	0.8582	0.8484	0.8493
	0.2	0.7699	0.8338	0.8581	0.8460	0.8469
	0.5	0.6964	0.8261	0.8574	0.8297	0.8310
	1	0.6128	0.8254	0.8650	0.8015	0.8028

Table 3.7 “True” and Estimated PC in Study 2

Test Length	Testlet Effect Variance	"True" PC	LL	LL_Strat	LEE	HH
36MC + 4FR	0	0.7632	0.7473	0.7628	0.7598	0.7581
	0.2	0.7705	0.7522	0.7714	0.7573	0.7560
	0.5	0.7662	0.7433	0.7722	0.7334	0.7361
	1	0.7783	0.7512	0.7858	0.7175	0.7180
28MC + 8FR	0	0.7925	0.7765	0.8005	0.7903	0.7892
	0.2	0.7961	0.7685	0.8008	0.7871	0.7881
	0.5	0.8000	0.7619	0.8020	0.7694	0.7705
	1	0.8118	0.7643	0.8152	0.7427	0.7456

Table 3.8 “True” and Estimated Kappa in Study 2

Test Length	Testlet Effect Variance	"True" Kappa	LL	LL_Strat	LEE	HH
36MC + 4FR	0	0.6340	0.6082	0.6323	0.6286	0.6260
	0.2	0.6383	0.6090	0.6394	0.6190	0.6172
	0.5	0.6320	0.5948	0.6406	0.5870	0.5911
	1	0.6399	0.5968	0.6527	0.5551	0.5555
28MC + 8FR	0	0.6701	0.6455	0.6834	0.6658	0.6640
	0.2	0.6780	0.6343	0.6853	0.6623	0.6643
	0.5	0.6872	0.6257	0.6892	0.6413	0.6431
	1	0.7004	0.6243	0.7053	0.6016	0.6058

Table 3.9 Bias of PA Estimates in Study 2

Test Length	Testlet Effect Variance	LL	LL_Strat	LEE	HH
36MC + 4FR	0	-0.0084	0.0039	-0.0005	0.0005
	0.2	***0.0667***	***0.0821***	***0.0680***	***0.0674***
	0.5	***0.1303***	***0.1549***	***0.1180***	***0.1179***
	1	***0.1986***	***0.2275***	***0.1607***	***0.1613***
28MC + 8FR	0	-0.0085	0.0092	-0.0005	0.0004
	0.2	***0.0639***	***0.0882***	***0.0761***	***0.0771***
	0.5	***0.1297***	***0.1610***	***0.1333***	***0.1346***
	1	***0.2126***	***0.2522***	***0.1887***	***0.1900***

Table 3.10 Bias of PC Estimates in Study 2

Test Length	Testlet Effect Variance	LL	LL_Strat	LEE	HH
36MC + 4FR	0	-0.0159	-0.0003	-0.0034	-0.0051
	0.2	-0.0183	0.0009	-0.0132	-0.0145
	0.5	-0.0229	0.0060	-0.0328	-0.0301
	1	-0.0271	0.0075	***-0.0608***	***-0.0603***
28MC + 8FR	0	-0.0159	0.0080	-0.0022	-0.0032
	0.2	-0.0276	0.0047	-0.0090	-0.0080
	0.5	-0.0381	0.0020	-0.0306	-0.0295
	1	-0.0475	0.0034	***-0.0691***	***-0.0662***

Table 3.11 Bias of Kappa Estimates in Study 2

Test Length	Testlet Effect Variance	LL	LL_Strat	LEE	HH
36MC + 4FR	0	-0.0258	-0.0017	-0.0054	-0.0080
	0.2	-0.0293	0.0010	-0.0193	-0.0212
	0.5	-0.0372	0.0086	-0.0450	-0.0408
	1	-0.0431	0.0128	***-0.0849***	***-0.0845***
28MC + 8FR	0	-0.0246	0.0133	-0.0043	-0.0061
	0.2	-0.0437	0.0072	-0.0157	-0.0137
	0.5	***-0.0614***	0.0020	-0.0458	-0.0441
	1	***-0.0762***	0.0048	***-0.0988***	***-0.0947***

Table 3.12 “True” and Estimated DC/DA Indices in Study 3

Index	Fitted Models	Truth	LL	LL_Strat	LEE	HH
PA	1PL/PCM	0.8054	0.8735	0.8822	0.8767	0.8762
	3PL/GRM	0.8273	0.8151	0.8275	0.8246	0.8267
PC	1PL/PCM	0.8284	0.8202	0.8327	0.8284	0.8262
	3PL/GRM	0.7638	0.7458	0.7614	0.7603	0.7619
Kappa	1PL/PCM	0.6812	0.6675	0.6904	0.6836	0.6799
	3PL/GRM	0.6302	0.6062	0.6305	0.6290	0.6313

Table 3.13 Bias of DC/DA Estimates in Study 3

Index	Fitted Models	LL	LL_Strat	LEE	HH
PA	1PL/PCM	***0.0681***	***0.0768***	***0.0713***	***0.0708***
	3PL/GRM	-0.0122	0.0002	-0.0027	-0.0006
PC	1PL/PCM	-0.0082	0.0042	0.0000	-0.0022
	3PL/GRM	-0.0180	-0.0024	-0.0035	-0.0019
Kappa	1PL/PCM	-0.0137	0.0092	0.0024	-0.0013
	3PL/GRM	-0.0240	0.0003	-0.0012	0.0011

Table 3.14 Weights of MC and FR Item Score in Composite Score in Study 4

Test Length	MC	FR
10	6.6667	10
20	3.3333	5
40	1.6667	2.5
80	0.8333	1.25

Table 3.15 “True” and Estimated PA in Study 4

Test Length	Metric	"True" PA	LL	LL_Strat	LEE	HH
10	Theta	0.7401	NA		NA	0.7597
	Raw	0.6805	0.6611		0.6599	0.6602
	Composite	0.6896	0.6453		NA	0.6808
20	Theta	0.8019	NA	NA	NA	0.8147
	Raw	0.7543	0.7393	0.7646	0.7438	0.7430
	Composite	0.7708	0.7415	0.7930	NA	0.7645
40	Theta	0.8478	NA	NA	NA	0.8479
	Raw	0.8273	0.8151	0.8275	0.8246	0.8267
	Composite	0.8311	0.8186	0.8447	NA	0.8290
80	Theta	0.8897	NA	NA	NA	0.8882
	Raw	0.8778	0.8722	0.8825	0.8757	0.8764
	Composite	0.8788	0.8722	0.8908	NA	0.8767

Table 3.16 “True” and Estimated PC in Study 4

Test Length	Metric	"True"PC	LL	LL_Strat	LEE	HH
10	Theta	0.6869	NA		NA	0.6678
	Raw	0.6246	0.5852		0.6142	0.6148
	Composite	0.5872	0.5243		NA	0.5878
20	Theta	0.7564	NA	NA	NA	0.7507
	Raw	0.6802	0.6565	0.6857	0.6695	0.6691
	Composite	0.6971	0.6450	0.7110	NA	0.6875
40	Theta	0.7930	NA	NA	NA	0.7922
	Raw	0.7638	0.7458	0.7614	0.7603	0.7619
	Composite	0.7664	0.7450	0.7811	NA	0.7647
80	Theta	0.8470	NA	NA	NA	0.8425
	Raw	0.8266	0.8197	0.8338	0.8265	0.8291
	Composite	0.8312	0.8197	0.8458	NA	0.8263

Table 3.17 “True” and Estimated Kappa in Study 4

Test Length	Metric	"True" Kappa	LL	LL_Strat	LEE	HH
10	Theta	0.4999	NA		NA	0.4703
	Raw	0.4147	0.3566		0.3969	0.3986
	Composite	0.3917	0.2874		NA	0.3806
20	Theta	0.6084	NA	NA	NA	0.6012
	Raw	0.5118	0.4734	0.5184	0.4947	0.4941
	Composite	0.5250	0.4627	0.5629	NA	0.5081
40	Theta	0.6713	NA	NA	NA	0.6715
	Raw	0.6302	0.6062	0.6305	0.6290	0.6313
	Composite	0.6345	0.6109	0.6660	NA	0.6316
80	Theta	0.7577	NA	NA	NA	0.7515
	Raw	0.7266	0.7153	0.7377	0.7258	0.7298
	Composite	0.7357	0.7230	0.7631	NA	0.7279

Table 3.18 Bias of PA Estimates in Study 4

Test Length	Metric	LL	LL_Strat	LEE	HH
10	Theta	NA		NA	0.0196
	Raw	-0.0194		-0.0206	-0.0203
	Composite	-0.0444		NA	-0.0089
20	Theta	NA	NA	NA	0.0128
	Raw	-0.0149	0.0103	-0.0105	-0.0113
	Composite	-0.0293	0.0222	NA	-0.0063
40	Theta	NA	NA	NA	0.0001
	Raw	-0.0122	0.0002	-0.0027	-0.0006
	Composite	-0.0125	0.0136	NA	-0.0021
80	Theta	NA	NA	NA	-0.0015
	Raw	-0.0055	0.0047	-0.0021	-0.0014
	Composite	-0.0067	0.0119	NA	-0.0021

Table 3.19 Bias of PC Estimates in Study 4

Test Length	Metric	LL	LL_Strat	LEE	HH
10	Theta	NA		NA	-0.0191
	Raw	-0.0395		-0.0104	-0.0098
	Composite	***-0.0629***		NA	0.0006
20	Theta	NA	NA	NA	-0.0057
	Raw	-0.0238	0.0054	-0.0107	-0.0112
	Composite	***-0.0521***	0.0140	NA	-0.0095
40	Theta	NA	NA	NA	-0.0009
	Raw	-0.0180	-0.0024	-0.0035	-0.0019
	Composite	-0.0214	0.0146	NA	-0.0017
80	Theta	NA	NA	NA	-0.0046
	Raw	-0.0069	0.0073	0.0000	0.0025
	Composite	-0.0115	0.0146	NA	-0.0048

Table 3.20 Bias of Kappa Estimates in Study 4

Test Length	Metric	LL	LL_Strat	LEE	HH
10	Theta	NA		NA	-0.0296
	Raw	***-0.0581***		-0.0178	-0.0161
	Composite	***-0.1044***		NA	-0.0112
20	Theta	NA	NA	NA	-0.0072
	Raw	-0.0384	0.0066	-0.0171	-0.0178
	Composite	***-0.0623***	0.0379	NA	-0.0169
40	Theta	NA	NA	NA	0.0003
	Raw	-0.0239	0.0003	-0.0011	0.0011
	Composite	-0.0236	0.0315	NA	-0.0028
80	Theta	NA	NA	NA	-0.0062
	Raw	-0.0113	0.0110	-0.0009	0.0031
	Composite	-0.0127	0.0275	NA	-0.0078

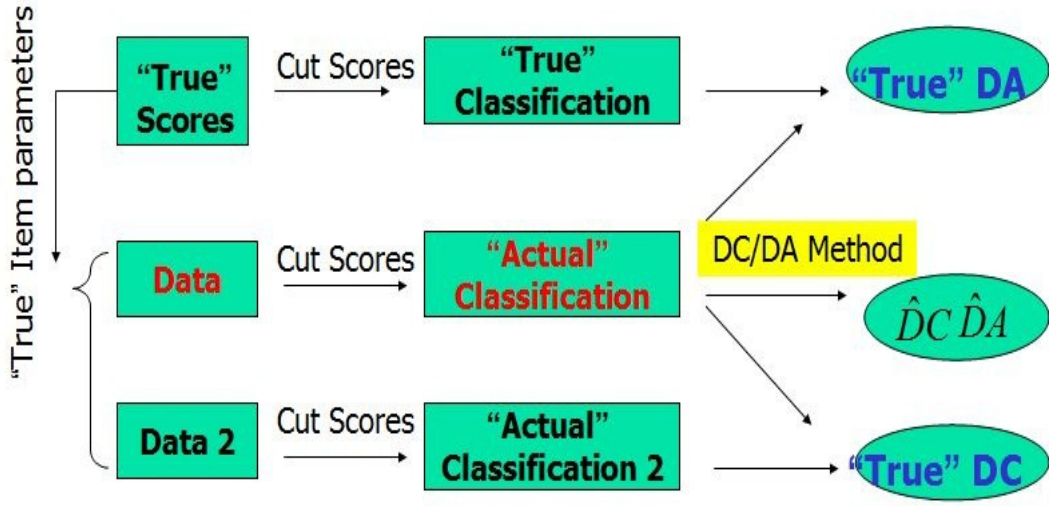


Figure 3.1 Calculations of "True" DC/DA Indices

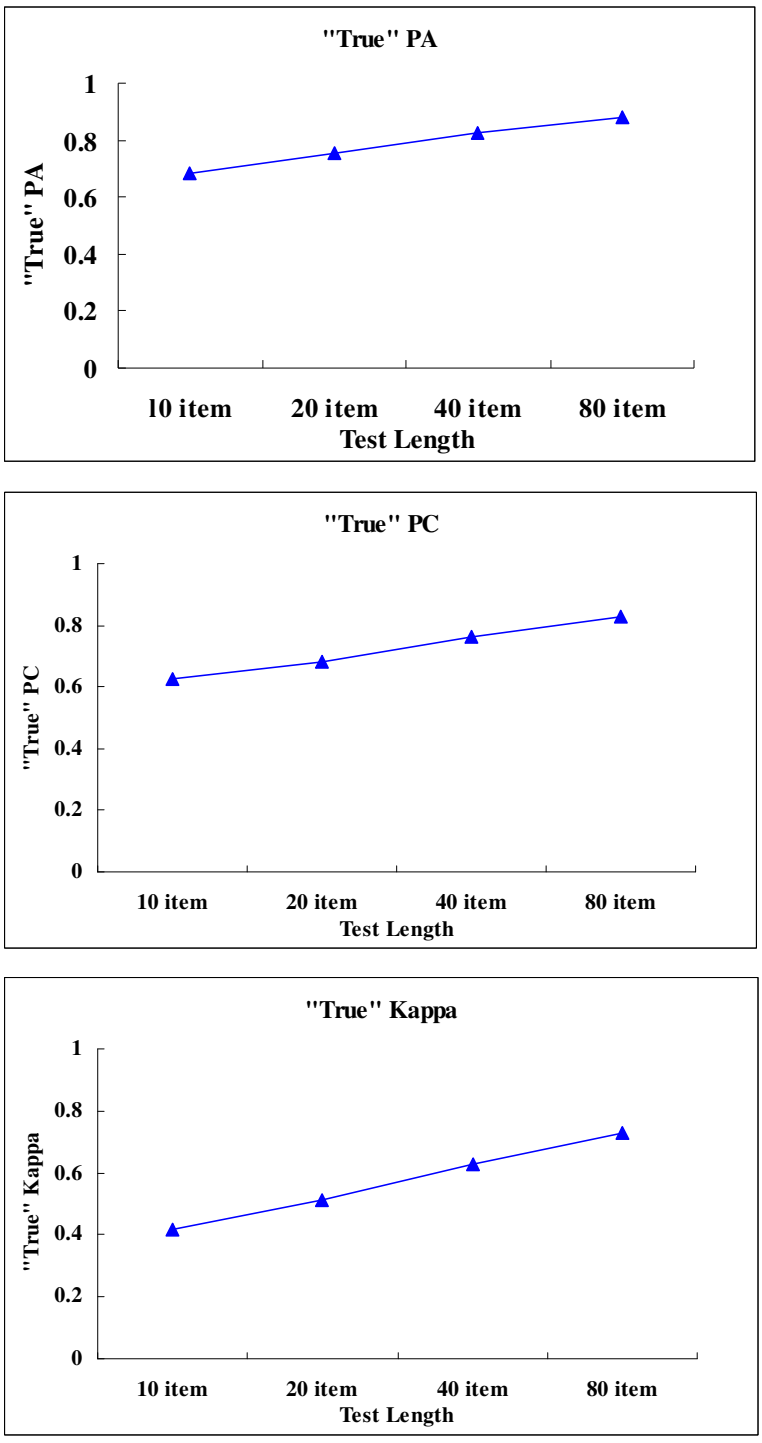


Figure 3.2 “True” DC/DA Indices in Study 1

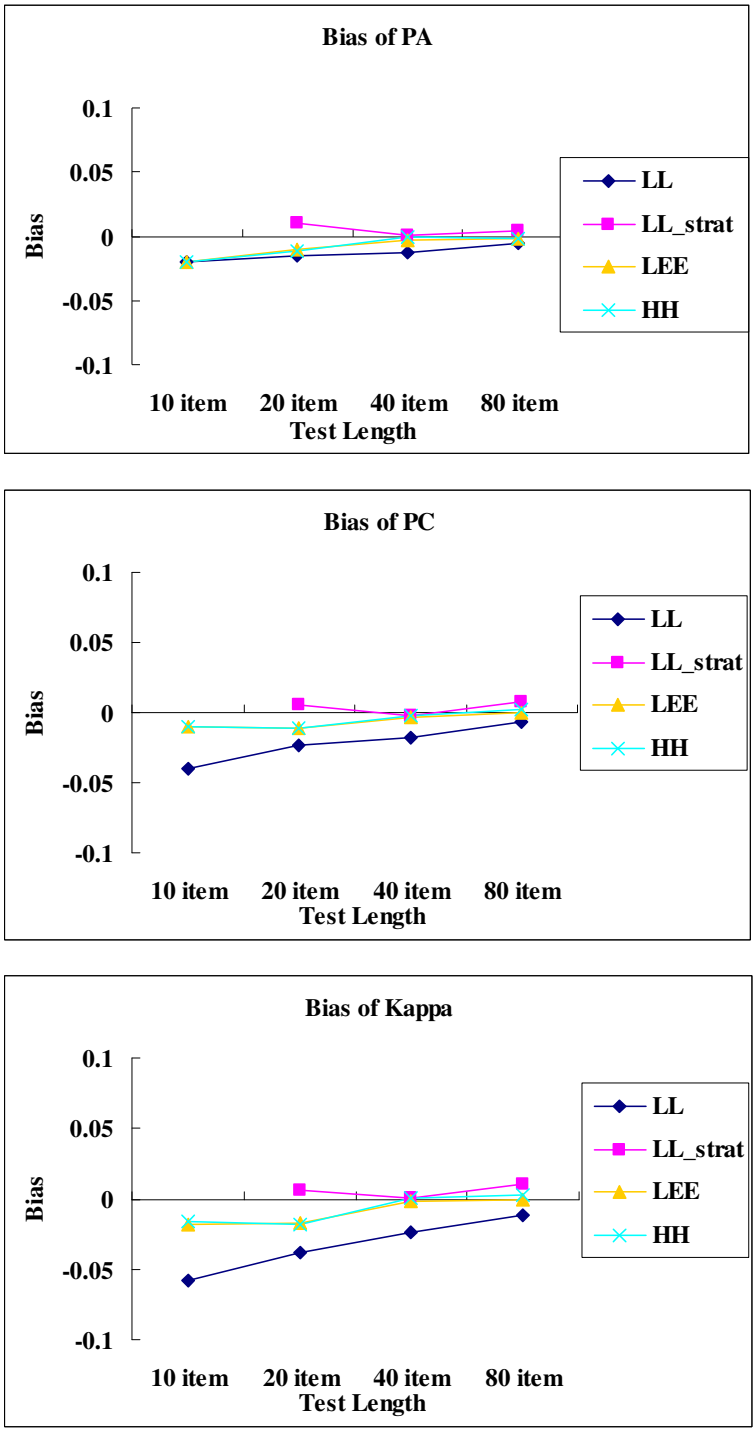


Figure 3.3 Bias of DC/DA Estimates in Study 1

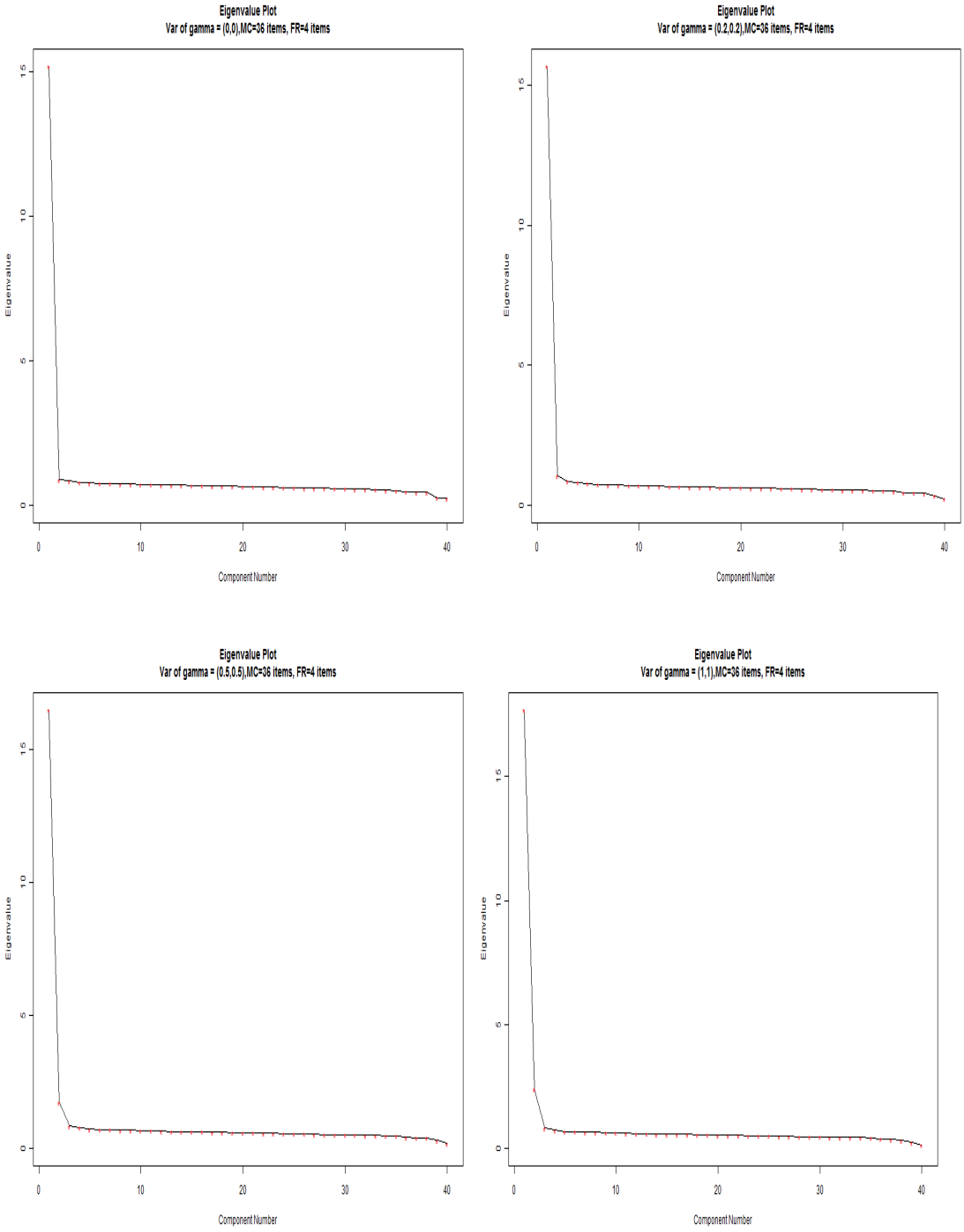


Figure 3.4 Eigenvalues of Tests with 36 MC and 4 FR (Variance of Gamma = 0, 0.2, 0.5, 1, from top left to bottom right)

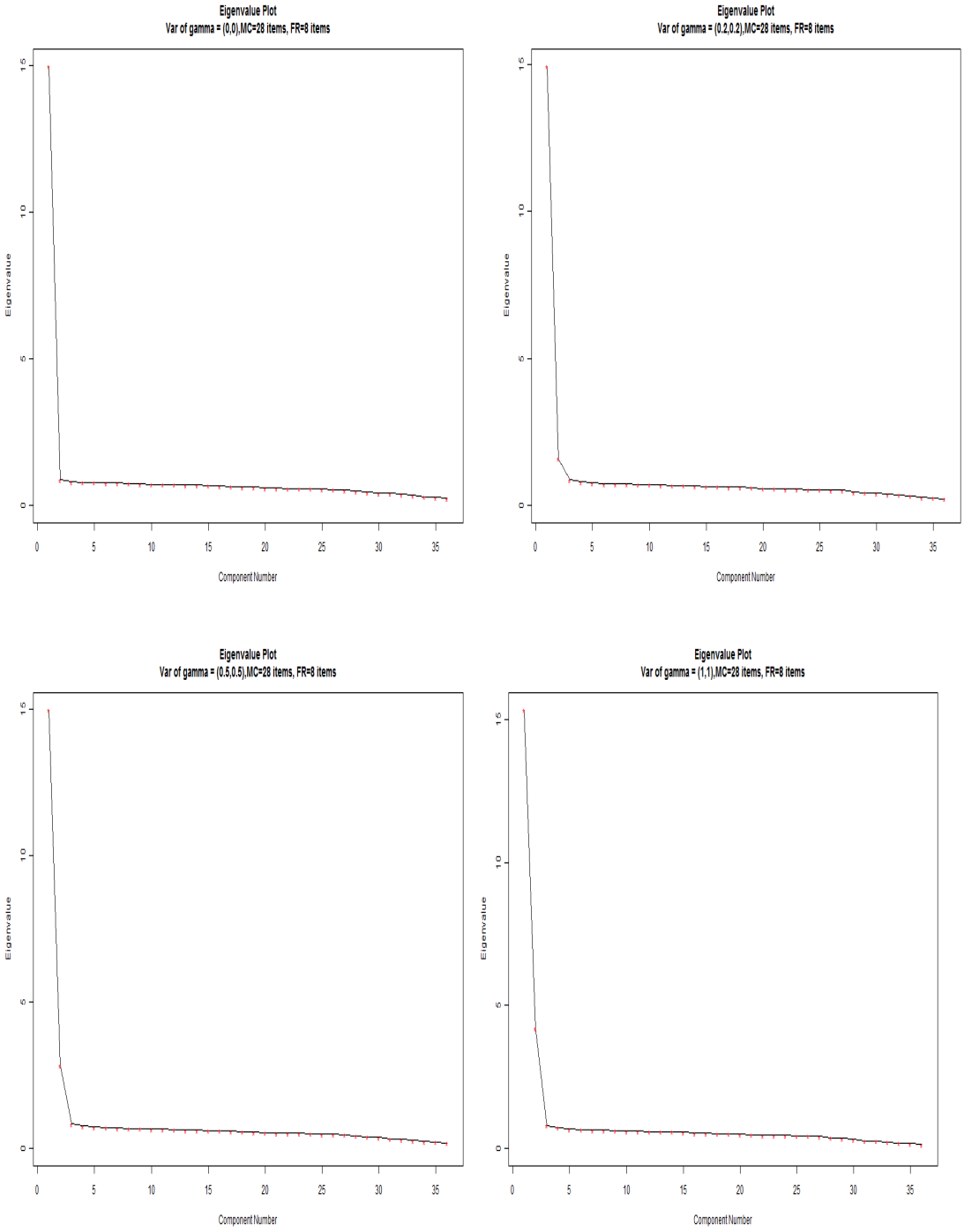


Figure 3.5 Eigenvalues of Tests with 28 MC and 8 FR (Variance of Gamma = 0, 0.2, 0.5, 1, from top left to bottom right)

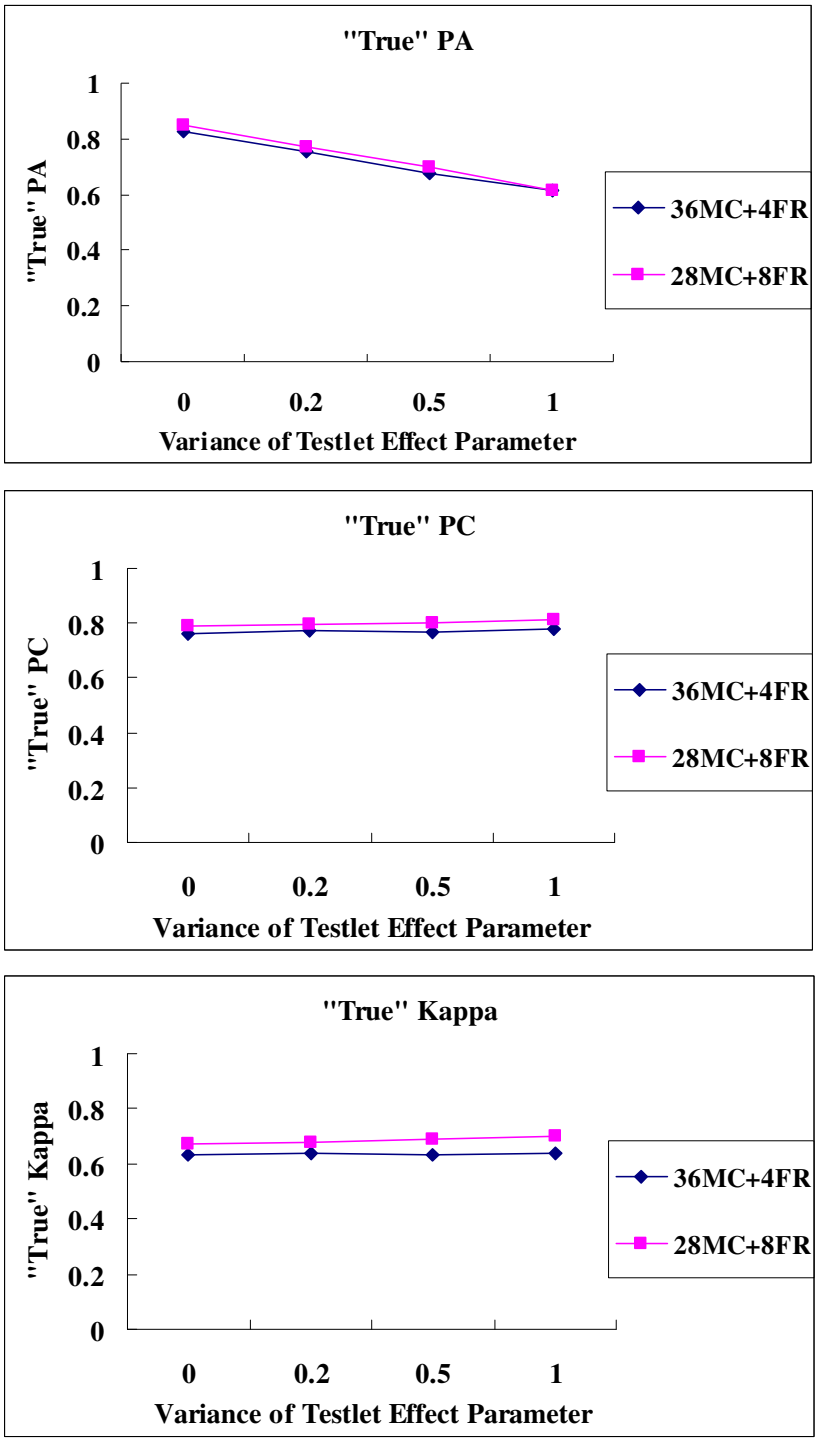
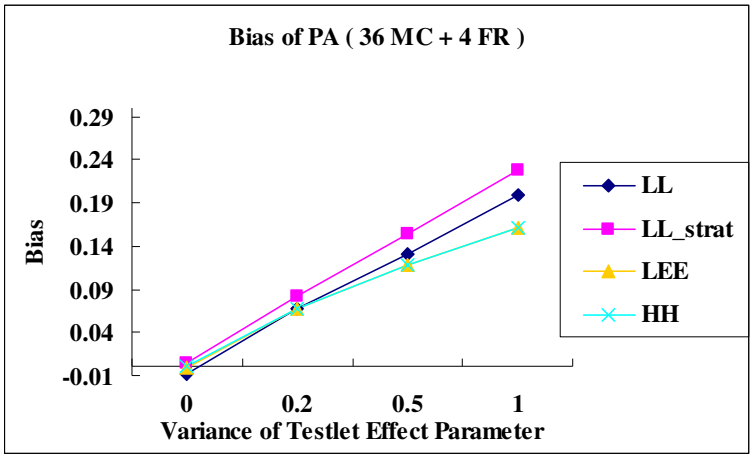
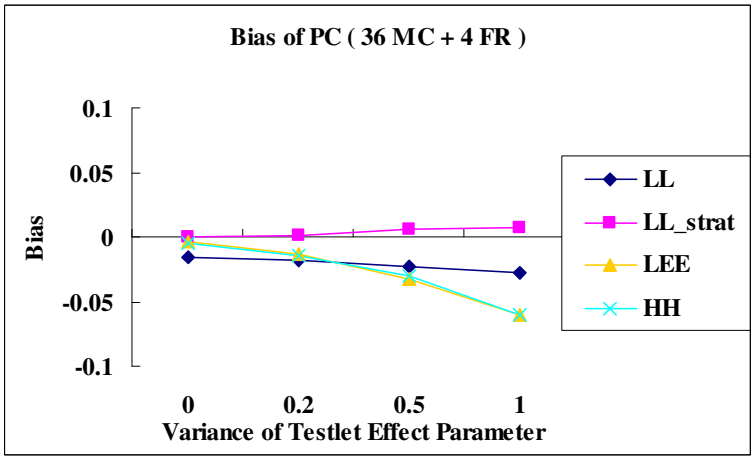


Figure 3.6 “True” DC/DA Indices of Different Conditions in Study 2





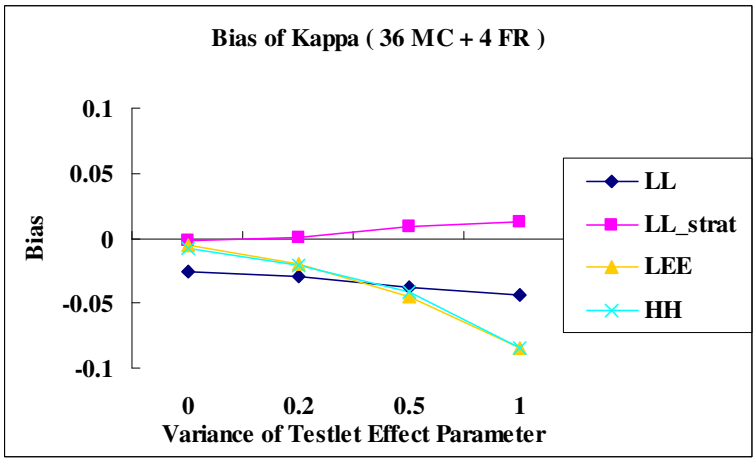
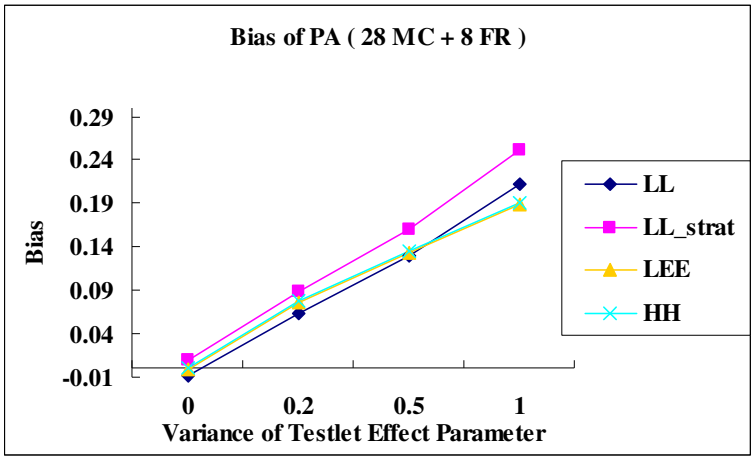
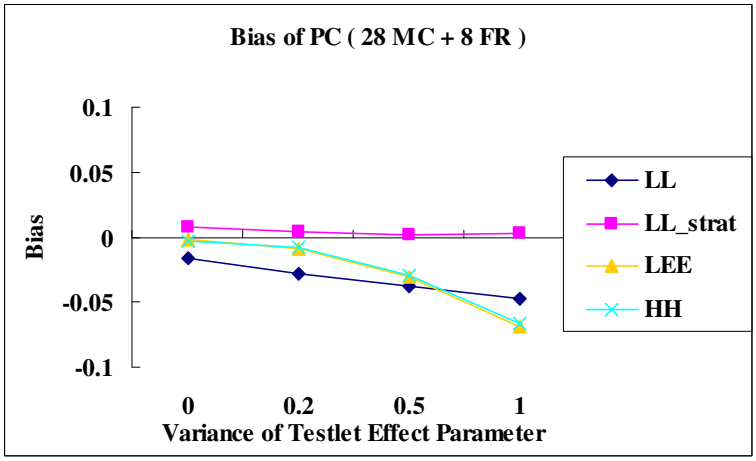


Figure 3.7 Bias of DC/DA Estimates for 36 MC + 4 FR in Study 2





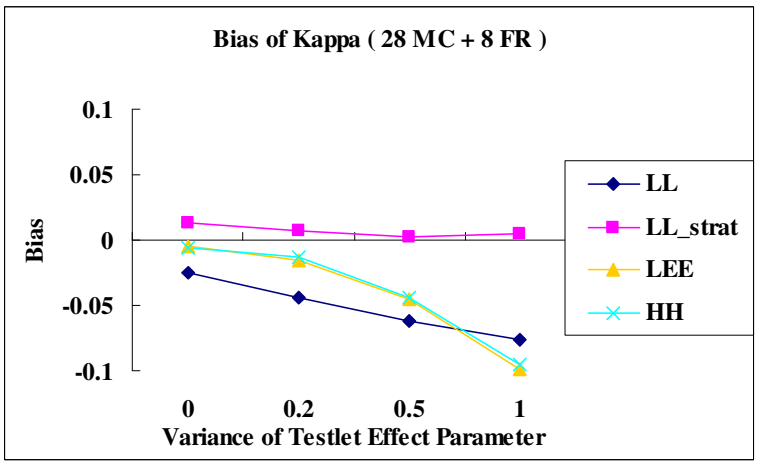


Figure 3.8 Bias of DC/DA Estimates for 28 MC + 8 FR in Study 2

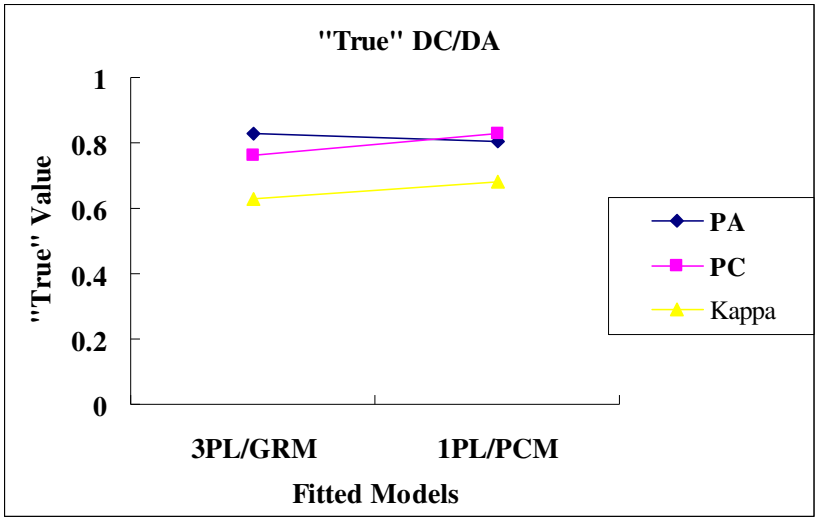


Figure 3.9 “True” DC/DA Index of Fitting Different IRT models in Study 3

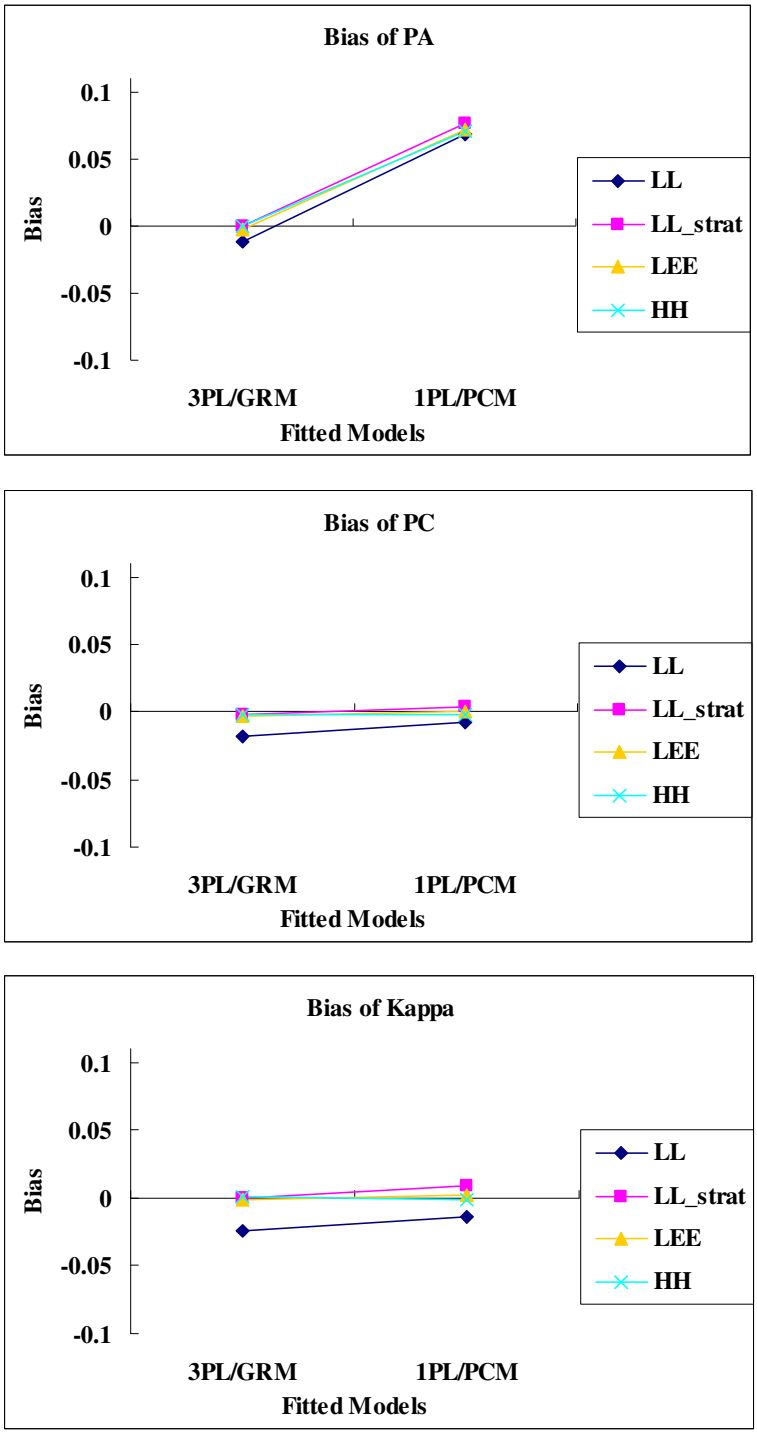


Figure 3.10 Bias of DC/DA Indices in Study 3

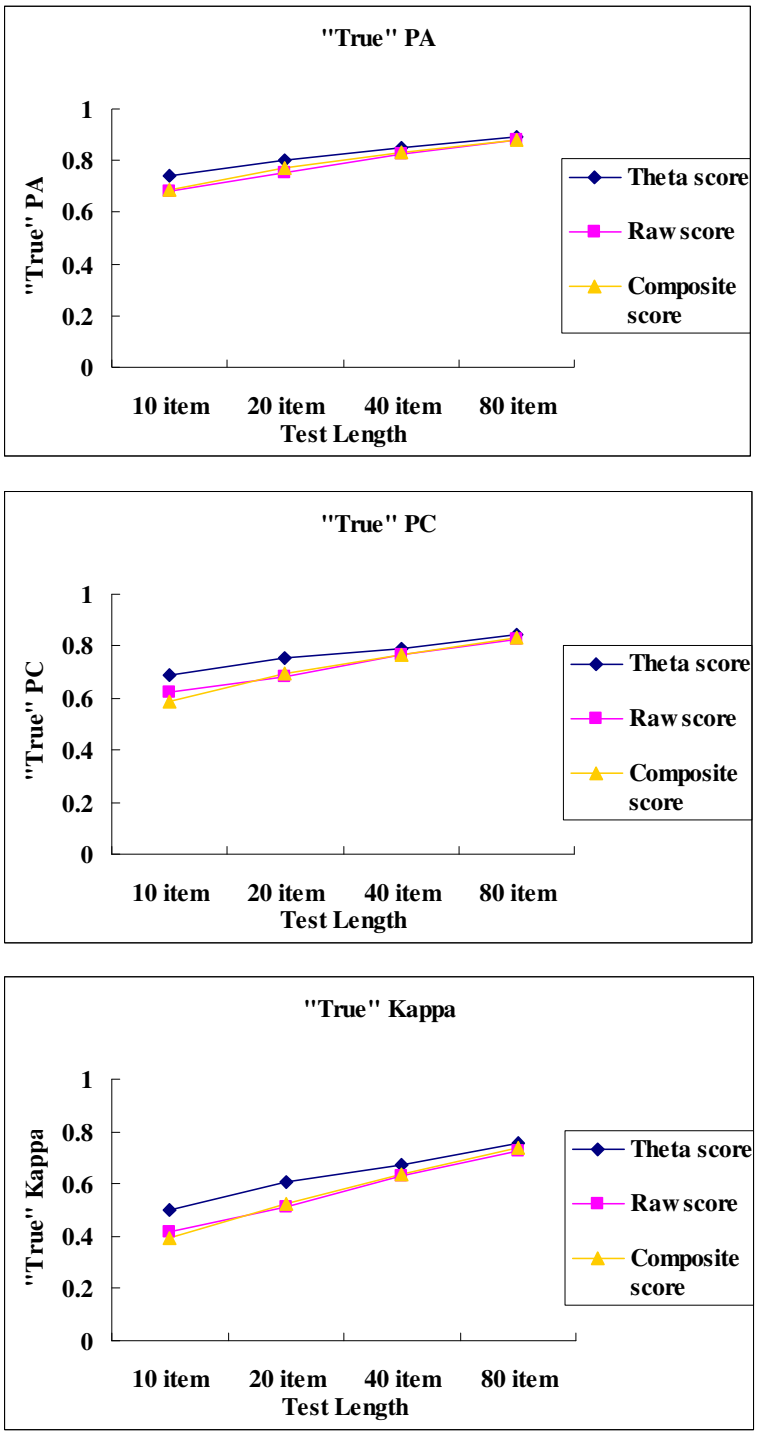
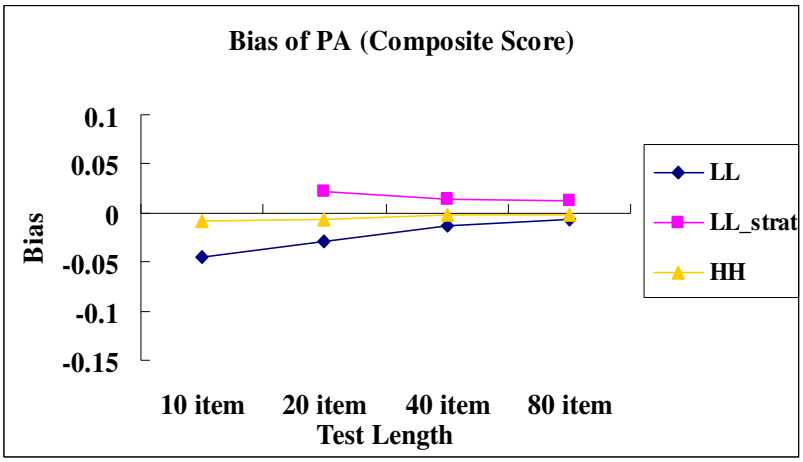
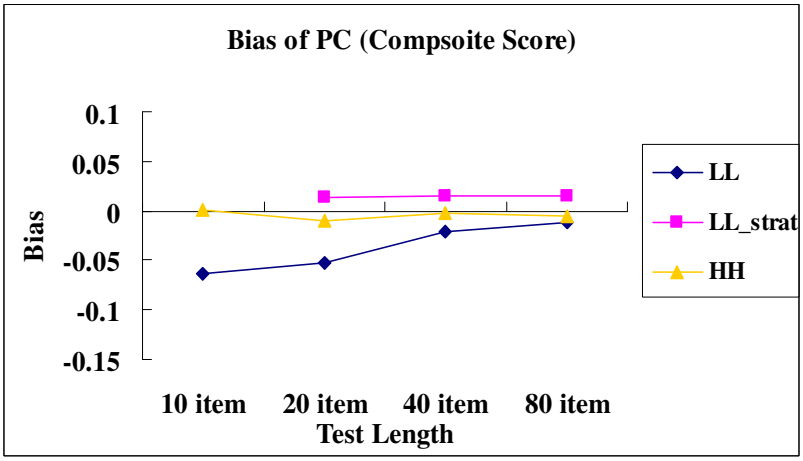


Figure 3.11 “True” DC/DA Estimates in Study 4





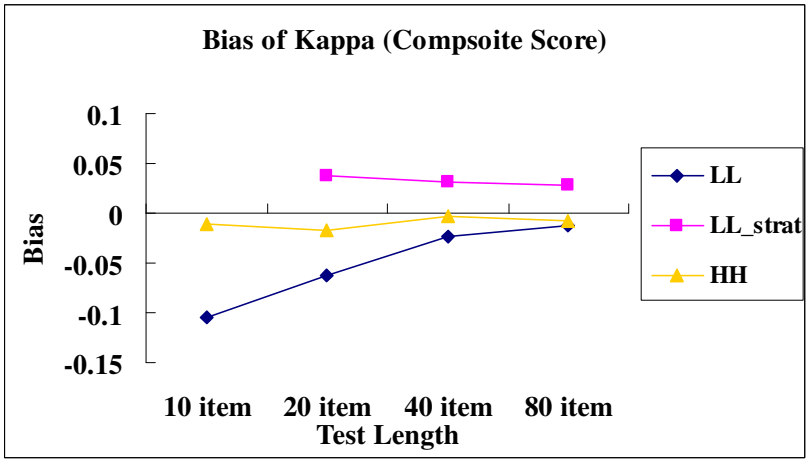
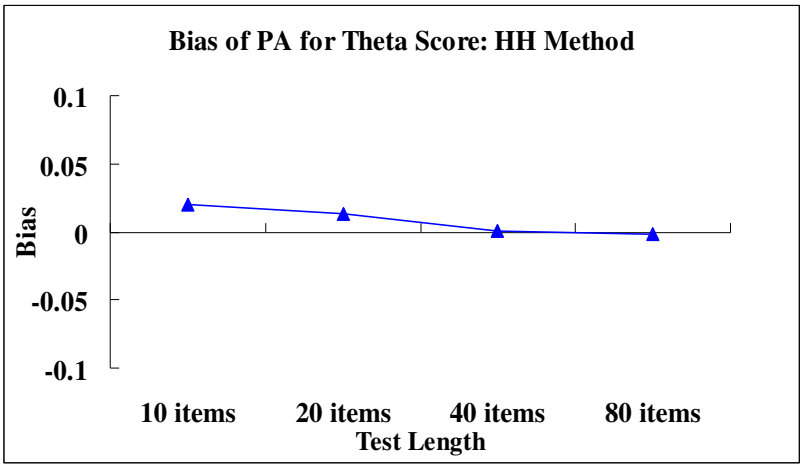
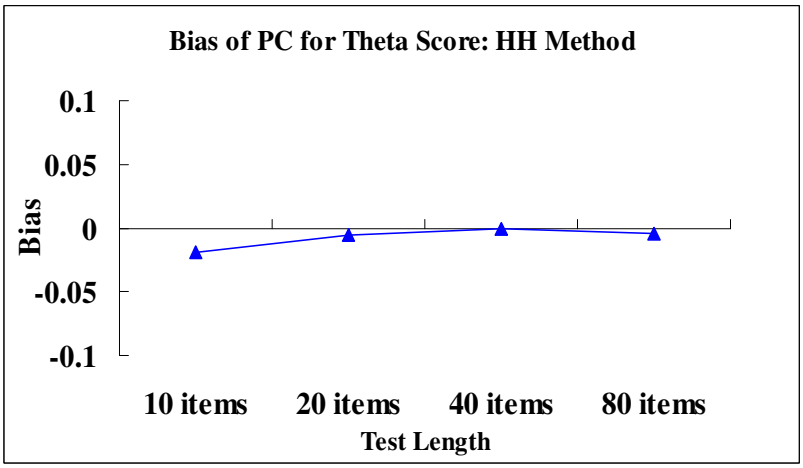


Figure 3.12 Bias of DC/DA Estimates on Composite Score Metric in Study 4





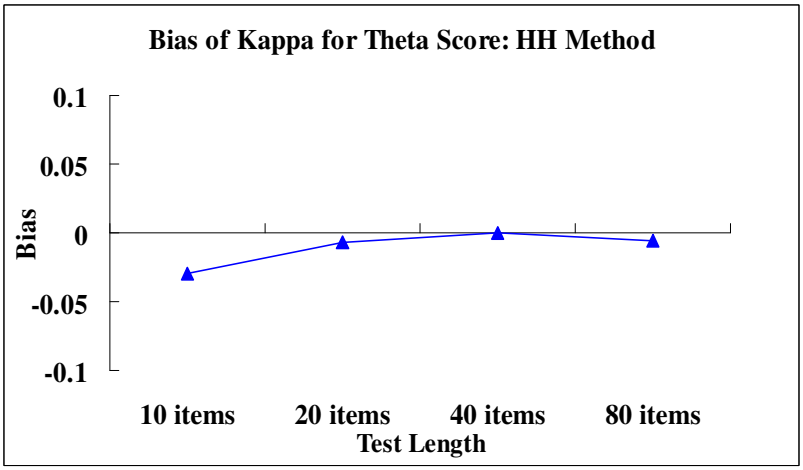


Figure 3.13 Bias of DC/DA Indices of HH Method on Theta Score Metric in Study 4

CHAPTER 4

REAL DATA STUDY

4.1. Method

4.1.1. Data

The real data study has the merits of assessing the performance of selected methods in real test conditions and truly reflecting the measurement errors without giving any advantages to certain models or assumptions. Therefore, the real data was used to evaluate the selected methods as a supplementary approach in addition to the simulation studies. Of course, while the results can be compared, truth is not known with real data and so when results are substantially different, it may be difficult to know which results are the most accurate. If the results are close, at least it is known that choice of method would be inconsequential on the results.

Ideally a pair of parallel forms are needed so that the actual DC estimate can be compared to any single administration estimates. Nevertheless, the use of parallel forms data is not always available. Alternatively, a long test can be split into halves and the two half-tests can be treated as parallel forms to calculate the decision consistency index. Then, of course the single administration estimates would work with only one of the two halves of the longer test. This is the design, for example, that Livingston and Lewis used in their research and was used a lot in literature (see Huynh, 1976; Livingston & Lewis, 1995, etc.). The DC index observed from the two half-tests is compared with the DC index estimated by the methods using one of the two half-tests.

A large-scale standardized achievement test was selected for the real data study. The original test (the Advanced Placement biology exam in 2006) was administrated to a large group of examinees in 3 hours. It consisted of 98 MC questions (scored 0-1) and 4 FR questions (scored 0-10). It was selected for the purpose because it was long in test length and consisted of both dichotomous and polytomous items. The data had a large sample size (20,000 examinees drawn from the original 131,783 test takers) so that the item and person ability parameters were well estimated in the IRT framework.

4.1.2. "True" DC Indices

The original full-length test was divided into two half-tests which were treated as two parallel forms. The cut scores were computed and applied to the two half-tests independently. A contingency table was constructed. The percentage of examinees who were classified consistently into the same category over the two half-tests was calculated and treated as the "true" PC, and the "true" Kappa was computed accordingly.

To split the original test into two half-tests, each having 49 MC items (scored 0-1) and 2 FR items (scored 0-10), a few steps were checked to make sure they were as parallel as possible, in terms of both item- and test-level statistics. The mean and standard deviation of the item parameters were compared in Table 4.1. Figure 4.1 plotted the raw score distributions for the full-length test and two half-tests. The reliability estimates of half-tests were check as well. The plots and numbers indicated that the two half-tests were quite comparable to each other.

The original full-length test classified the examinees into five grades. The observed percentages of examinees fallen into each of the five categories were 15.6%, 23.3%, 21.2%, 20.3%, and 19.6%, from grade 1 (the lowest) to grade 5 (the highest). The percentages were adopted to compute the cut scores for the two half-tests in the real data study. The cut scores were defined in the way that the same percentages of examinees were classified into each of the five categories based on their half-test scores. (This process is equivalent to what is called an "equipercentile equating" of the cut scores on the two halves of the test.)

Table 4.2 displayed the cut scores applied to the half-tests on raw score and composite score scale dividing the examinees into about the same percentages as specified above. The weights used in calculating the composite score were described in details in the following section.

Apply the cut scores to the half-tests independently and the contingency table was obtained. When applying the four cut scores simultaneously, the percentage of examinees who were classified consistently into the same category was calculated as the "true" PC. The "true" Kappa was calculated accordingly. In addition, the "true" PC and Kappa when each of the four cut scores was applied separately were calculated too. Table 4.4 summarized the "true" PC and Kappa for applying the cut scores both simultaneously (denoted as "All Cuts") and independently (denoted as "Cut1", "Cut2", "Cut3", and "Cut4"), and on the raw score and composite score metrics.

4.1.3. Factors Investigated

4.1.3.1. Reliability Estimate

To investigate the impacts of choice of reliability estimate on DC estimate of LL method using real data, three options of reliability estimates were considered (a) the standard Cronbach's alpha coefficient, (b) the stratified alpha coefficient, and (c) the correlation between the scores of two half-tests. The reliability estimates for the tests of different choices were summarized in Table 4.3. The variations based on the LL method were denoted as $LL_{Cronbach}$, LL_{strat} and LL_{corr}, separately.

4.1.3.2. Competing IRT Models

Different from the simulation studies where the true models were known, the true models were unknown in the real data study. IRT-based methods by fitting competing IRT models were used to the real data, and their DC/DA estimates were compared. The assumption of unidimensionality was checked using the principal component analysis (PCA) prior to the IRT calibration. The eigenvalue plot in Figure 4.2 suggested the original full test was unidimensional.

Three sets of competing IRT models used for the LEE and HH methods were (1) 1PL/PCM, (2) 2PL/GRM, and (3) 3PL/GRM, the model before the slash was for dichotomous items while the model after was for polytomous items.

4.1.3.3. Scoring Metric

Two scoring metrics were used for the real data: the raw score and the composite score. The original full-length test was scored by using a composite score given by

Composite score = 1.2245(MC raw score) + 2 (FR raw score)

where 1.2245 and 2 were the weights applied to each score point for the MC and FR items separately. Remembering that there were 98 MC items (scored 0-1) and 4 FR items (scored 0-10), the contributions by weighted MC and FR score to the composite score were 60% and 40%, separately. The scoring formula was adopted for the half-tests. This condition was not studied for the LEE method due to the reason explained in the previous chapter.

4.1.3.4. Summary of Conditions

In summary, there were nine variations of methods studied on raw score scale: Three for the LL method, three for the LEE method, and three for the HH method. And there were six variations of methods studied on composite score: Three for the LL method and three for the HH method. In total, there were 15 conditions included in the real data study to check the performance of the LL_based and IRT_based methods using their different options.

4.2. Results

4.2.1. Raw Score

Table 4.5 and Table 4.6 displayed the PC and Kappa estimates for the variations of the LL method. Each variation was used for both the first and the second half-test. The tables showed that the estimates derived from two of the half-tests were very close to each other. The estimates from only the first half-test were used to plot for illustration. Figure 4.3 plotted the PC and Kappa estimates against the truth. The plots showed that the LL method with correlation of half-tests as the reliability estimate produced the

PC/Kappa estimates the closest to the "true" PC/Kappa indices (Note that this was not the reliability estimate used by Livingston and Lewis and is not the reliability estimate typically used in practice). This makes sense since the "true" PC/Kappa indices were calculated based on observation from the two half-tests. Besides, the LL method using the stratified alpha had good and accurate estimates too. The LL method using Cronbach's alpha under-estimated the indices by about 0.05. The under-estimation was persistent when the cut scores were applied in different ways (whether multiple cuts applied together or single cut applied separately).

Table 4.7 to Table 4.10 provided the PC and Kappa estimates for the variations of IRT-based methods. Again the estimates from both of the half-tests were very similar (differences on the third decimal) and the estimates from half-test 1 were plotted in Figure 4.4 for illustration. The plots showed that the LEE and HH methods using the 2PL/GRM and 3PL/GRM produced almost identical results with the "true" DC indices. The methods using the 1PL/PCM tended to over-estimate the DC indices by 0.03.

4.2.2. Composite Score

Table 4.11 to Table 4.14 provided the PC and Kappa estimates on the composite score scale for the LL and HH methods. Figure 4.5 plotted the estimates against the "true" indices on the composite score scale. The plots showed that the DC estimates on composite score scale had close pattern with the estimates on the raw score scale. The LL method using Cronbach's alpha persistently underestimated the indices on the composite score scale, and the difference was beyond 0.1 and larger than on the raw

score scale. The HH method with 1PL/PCM again overestimated the indices by around 0.05. All the other variations of methods resulted in accurate estimates.

Table 4.1 Mean and SD of Item Parameters in the Test

Test	a	b	c
Full-length Test	(1.29, 0.40)	(0.09, 1.29)	(0.18, 0.10)
Half-test 1	(1.31, 0.40)	(0.08, 1.37)	(0.19, 0.09)
Half-test 2	(1.28, 0.40)	(0.09, 1.21)	(0.17, 0.10)

Table 4.2 Cut Scores of Half-Tests

Score	Test	Cut 1	Cut 2	Cut 3	Cut 4
Raw Score	Half-Test 1	24	34	41	49
	Half-Test 2	24	34	41	48
Composite Score	Half-Test 1	31.71	47.27	58.86	70.86
	Half-Test 2	32.49	46.41	57.18	68.53

Table 4.3 Reliability Estimates of Different Choices

Test	Correlation	Cronbach's Alpha	Stratified Alpha
Full-Length Test	/	0.922	0.944
Half-Test 1	0.895	0.846	0.900
Half-Test 2	0.895	0.860	0.888

Table 4.4 "True" PC and Kappa Indices

Metric	Cut	PC	Kappa
Raw Score	All Cuts	0.574	0.466
	Cut 1	0.919	0.681
	Cut 2	0.868	0.721
	Cut 3	0.856	0.705
	Cut 4	0.888	0.666
Composite Score	All Cuts	0.558	0.445
	Cut 1	0.917	0.680
	Cut 2	0.858	0.702
	Cut 3	0.849	0.685
	Cut 4	0.884	0.634

Table 4.5 PC Estimates on Raw Score Metric: LL Method

Cut	Corr _Test1	Corr _Test2	Cronbach _Test1	Cronbach _Test2	Strat _Test1	Strat _Test2
All Cuts	0.575	0.578	0.515	0.533	0.584	0.569
Cut1	0.917	0.918	0.900	0.905	0.919	0.916
Cut2	0.867	0.866	0.840	0.846	0.871	0.862
Cut3	0.859	0.862	0.830	0.840	0.864	0.858
Cut4	0.889	0.889	0.866	0.871	0.892	0.886

Table 4.6 Kappa Estimates on Raw Score Metric: LL Method

Cut	Corr _Test1	Corr _Test2	Cronbach _Test1	Cronbach _Test2	Strat Test1	Strat _Test2
All Cuts	0.467	0.469	0.392	0.412	0.478	0.458
Cut1	0.681	0.671	0.614	0.622	0.689	0.662
Cut2	0.719	0.717	0.660	0.676	0.727	0.710
Cut3	0.711	0.715	0.652	0.669	0.720	0.706
Cut4	0.662	0.673	0.590	0.619	0.671	0.664

Table 4.7 PC Estimates on Raw Score Metric: LEE Method

Cut	1PL/PCM _Test1	1PL/PCM _Test2	2PL/GRM _Test1	2PL/GRM _Test2	3PL/GRM _Test1	3PL/GRM _Test2
All Cuts	0.611	0.611	0.575	0.573	0.577	0.576
Cut1	0.924	0.926	0.922	0.921	0.919	0.918
Cut2	0.881	0.880	0.866	0.866	0.864	0.865
Cut3	0.874	0.876	0.857	0.857	0.858	0.860
Cut4	0.902	0.899	0.891	0.887	0.896	0.891

Table 4.8 Kappa Estimates on Raw Score Metric: LEE Method

Cut	1PL/PCM _Test1	1PL/PCM _Test2	2PL/GRM _Test1	2PL/GRM _Test2	3PL/GRM _Test1	3PL/GRM _Test2
All Cuts	0.512	0.511	0.466	0.462	0.468	0.465
Cut1	0.718	0.710	0.683	0.666	0.674	0.658
Cut2	0.749	0.747	0.713	0.716	0.712	0.715
Cut3	0.743	0.744	0.706	0.705	0.707	0.710
Cut4	0.711	0.712	0.658	0.657	0.670	0.669

Table 4.9 PC Estimates on Raw Score Metric: HH Method

Cut	1PL/PCM _Test1	1PL/PCM _Test2	2PL/GRM _Test1	2PL/GRM _Test2	3PL/GRM _Test1	3PL/GRM _Test2
All Cuts	0.610	0.614	0.577	0.576	0.577	0.575
Cut1	0.925	0.925	0.920	0.924	0.919	0.917
Cut2	0.881	0.881	0.867	0.865	0.861	0.865
Cut3	0.876	0.878	0.857	0.857	0.860	0.862
Cut4	0.899	0.898	0.894	0.889	0.898	0.891

Table 4.10 Kappa Estimates on Raw Score Metric: HH Method

Cut	1PL/PCM _Test1	1PL/PCM _Test2	2PL/GRM _Test1	2PL/GRM _Test2	3PL/GRM _Test1	3PL/GRM _Test2
All Cuts	0.511	0.514	0.468	0.466	0.468	0.464
Cut1	0.720	0.710	0.672	0.671	0.676	0.654
Cut2	0.748	0.750	0.716	0.714	0.705	0.716
Cut3	0.747	0.749	0.707	0.705	0.711	0.714
Cut4	0.703	0.709	0.668	0.665	0.677	0.670

Table 4.11 PC Estimates on Composite Score Metric: LL Method

Cut	Corr _Test1	Corr _Test2	Cronbach _Test1	Cronbach _Test2	Strat Test1	Strat _Test2
All Cuts	0.560	0.559	0.463	0.479	0.570	0.538
Cut1	0.911	0.908	0.880	0.882	0.914	0.902
Cut2	0.861	0.858	0.815	0.817	0.865	0.847
Cut3	0.855	0.857	0.805	0.816	0.860	0.845
Cut4	0.887	0.891	0.844	0.860	0.890	0.884

Table 4.12 Kappa Estimates on Composite Score Metric: LL Method

Cut	Corr _Test1	Corr _Test2	Cronbach _Test1	Cronbach _Test2	Strat Test1	Strat _Test2
All Cuts	0.448	0.447	0.326	0.346	0.460	0.420
Cut1	0.660	0.643	0.542	0.544	0.671	0.620
Cut2	0.709	0.701	0.611	0.615	0.715	0.679
Cut3	0.697	0.701	0.594	0.617	0.707	0.678
Cut4	0.641	0.654	0.504	0.556	0.653	0.632

Table 4.13 PC Estimates on Composite Score Metric: HH Method

Cut	1PL/PCM _Test1	1PL/PCM _Test2	2PL/GRM _Test1	2PL/GRM _Test2	3PL/GRM _Test1	3PL/GRM _Test2
All Cuts	0.601	0.601	0.555	0.543	0.562	0.538
Cut1	0.917	0.919	0.913	0.910	0.917	0.907
Cut2	0.878	0.874	0.855	0.849	0.856	0.849
Cut3	0.872	0.874	0.848	0.847	0.848	0.844
Cut4	0.902	0.901	0.889	0.881	0.894	0.881

Table 4.14 Kappa Estimates on Composite Score Metric: HH Method

Cut	1PL/PCM _Test1	1PL/PCM _Test2	2PL/GRM _Test1	2PL/GRM _Test2	3PL/GRM _Test1	3PL/GRM _Test2
All Cuts	0.500	0.500	0.440	0.426	0.449	0.419
Cut1	0.702	0.703	0.655	0.644	0.674	0.636
Cut2	0.745	0.738	0.694	0.680	0.697	0.681
Cut3	0.733	0.739	0.681	0.681	0.679	0.674
Cut4	0.696	0.700	0.632	0.622	0.648	0.621

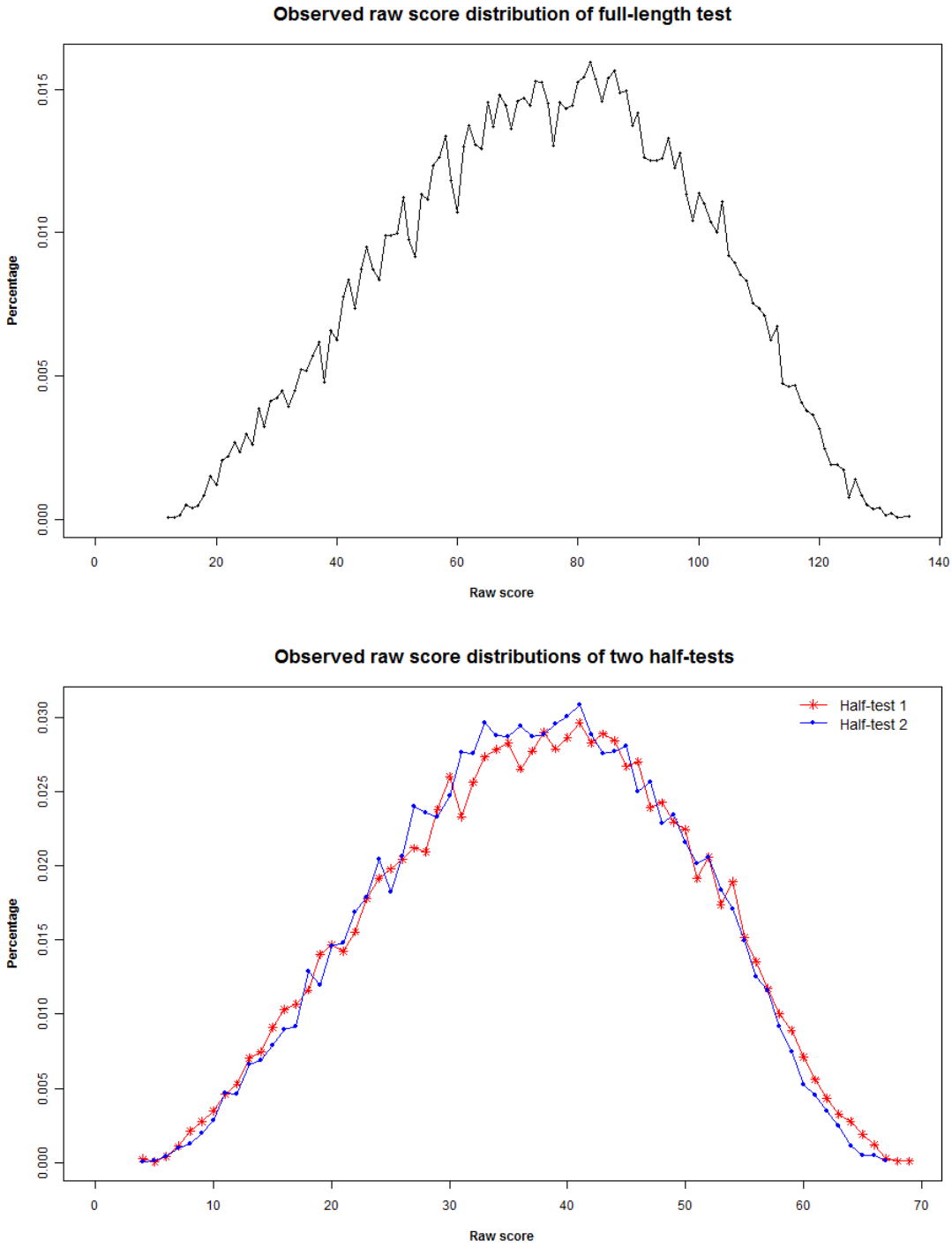


Figure 4. 1 Observed Raw Score Distributions of Full-length test and Two Half-tests

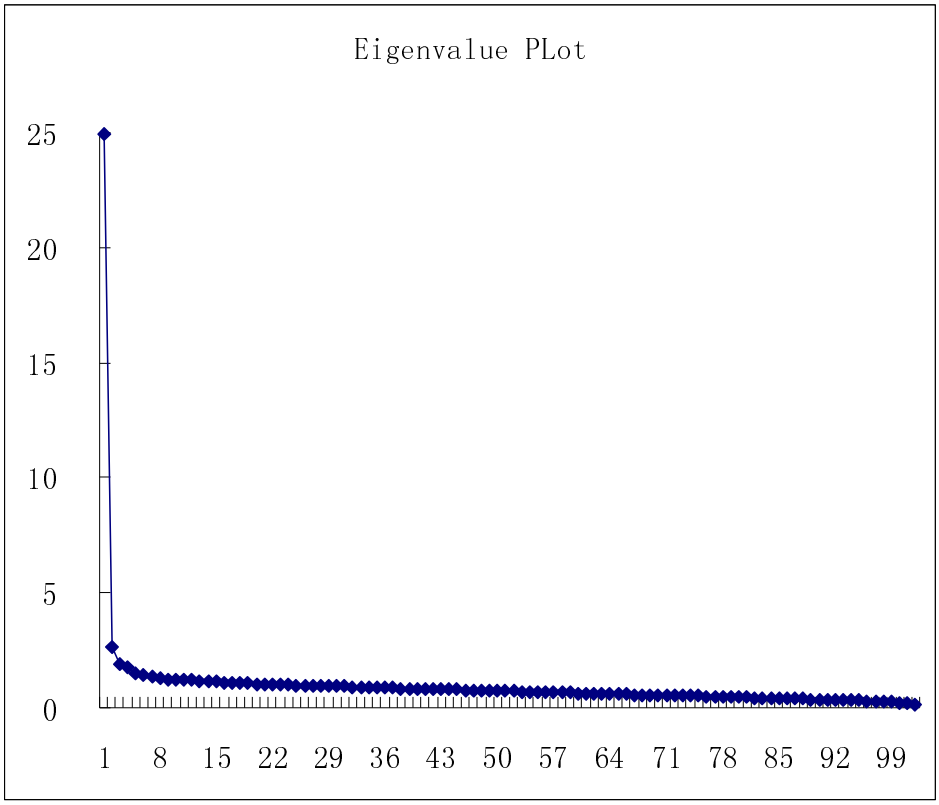
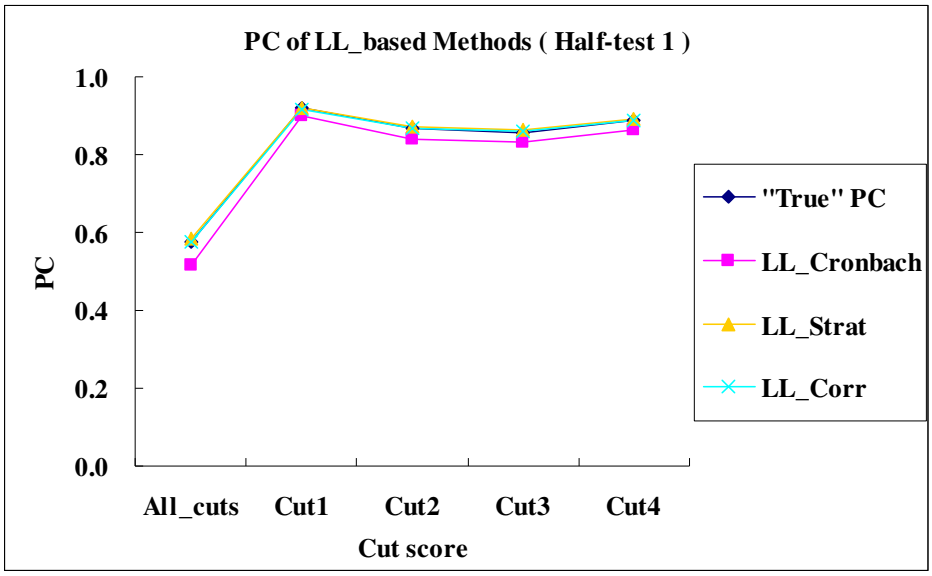


Figure 4.2 Eigenvalue Plot of Full-length Test



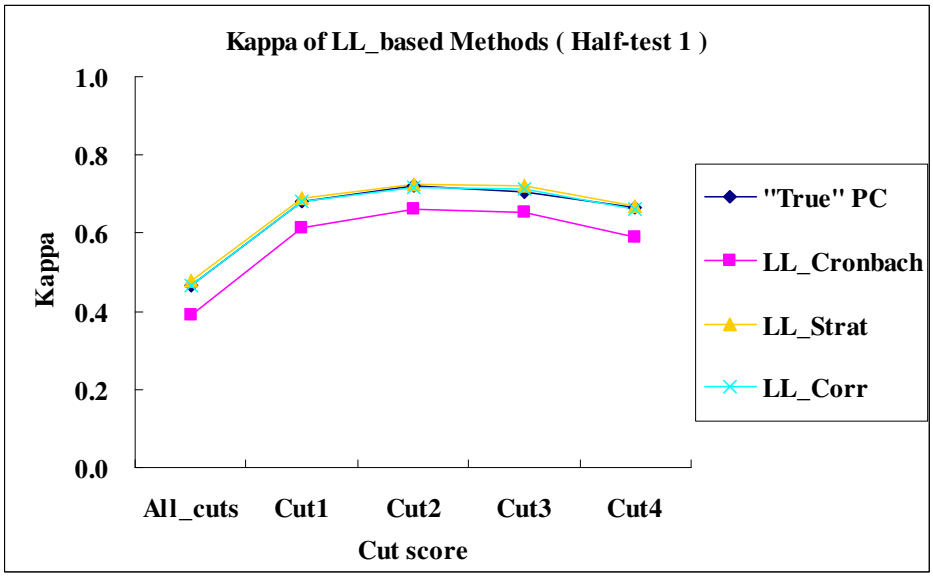
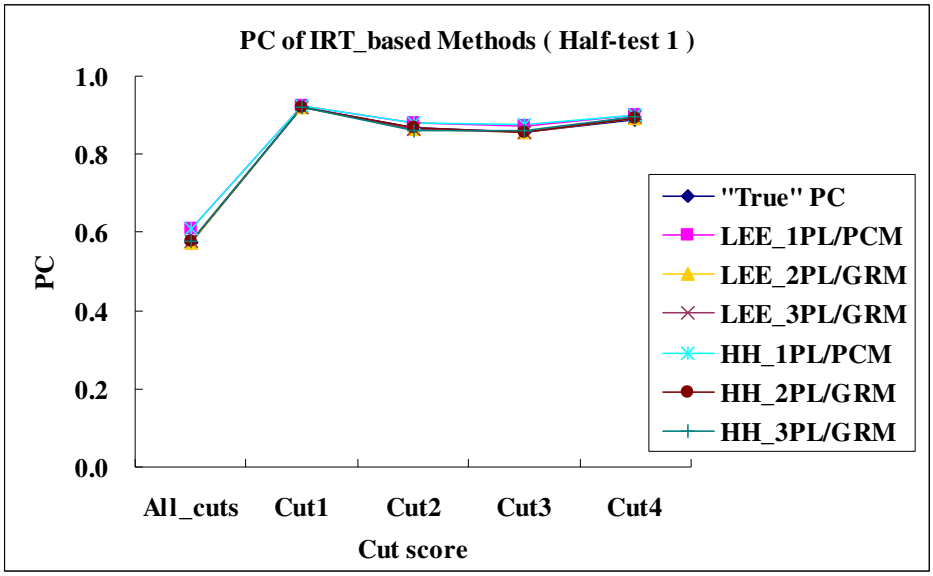


Figure 4.3 PC and Kappa Estimates of LL Method using Different Reliability Estimates



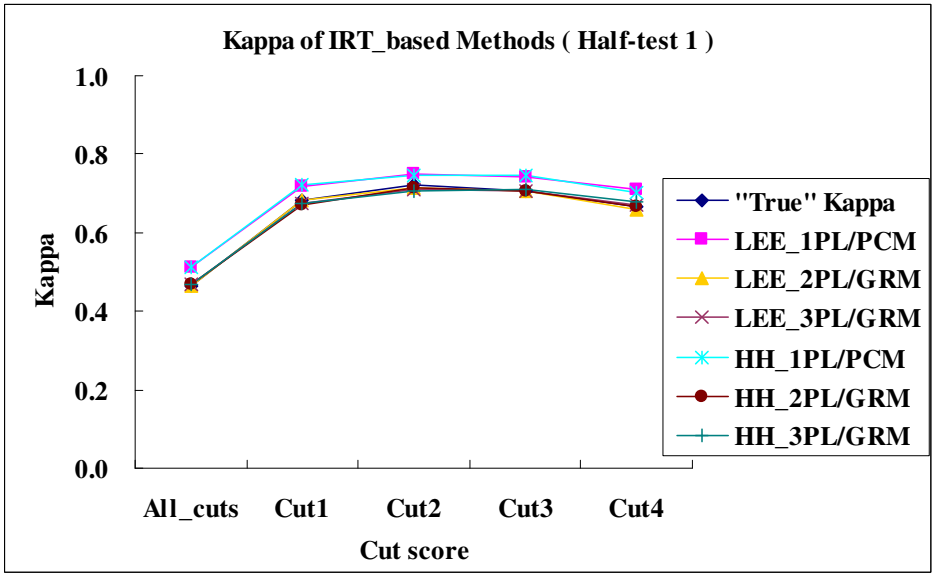
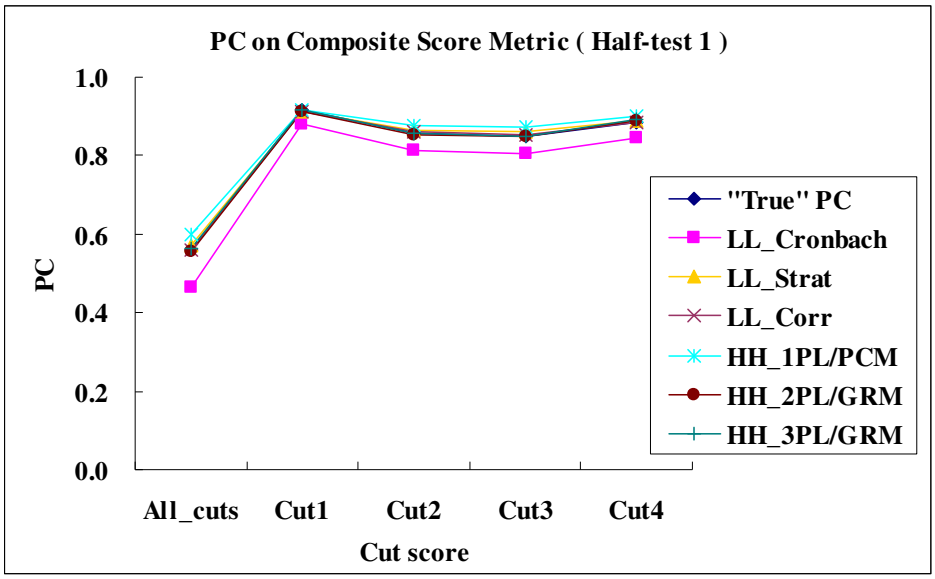


Figure 4.4 PC and Kappa Estimates of IRT-based Methods Fitting Different IRT Models



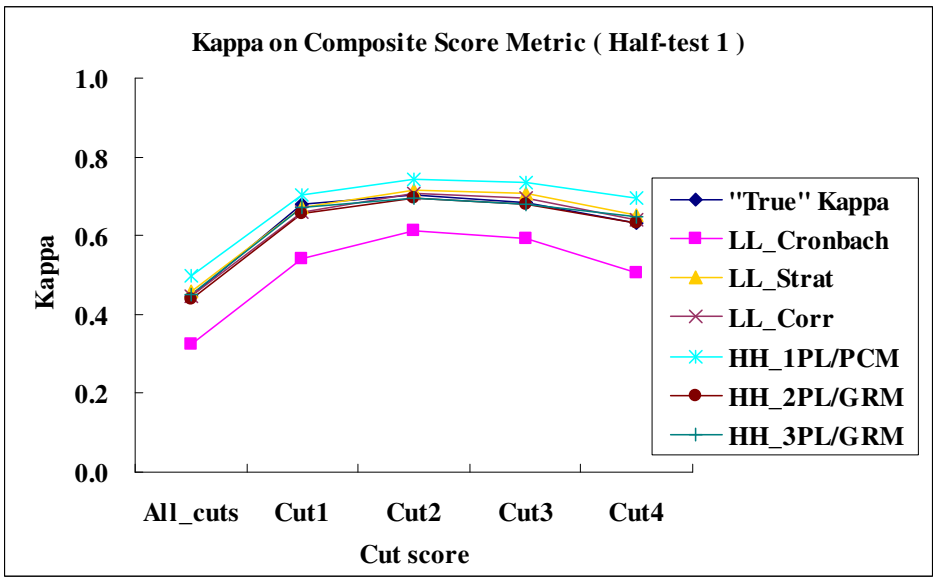


Figure 4.5 PC and Kappa Estimates of Different Methods on Composite Score Metric

CHAPTER 5

CONCLUSION AND DISCUSSION

5.1. Review of the Study

Four simulation studies and one empirical study were conducted in this dissertation to evaluate four variations of three major DC/DA methods: the LL, $LL_{strat,}$ LEE and HH methods. The robustness of these selected methods was evaluated against the factors of test length, local item dependency, model misfit and scoring metric on which the analyses are carried out.

The simulation studies were implemented in the IRT framework for two reasons: Firstly, IRT models provide good fit to educational test data and have been shown to be effective and useful in solving many problems in the educational measurement field,. Secondly, the IRT models are widely used so that the study would have more practical implications. The simulation studies were carried out as the primary approach because different conditions could be easily manipulated and the "true" DC/DA indices could be calculated. The absence of a meaningful criterion makes it nearly impossible to compare competing methods otherwise.

Study 1 looked at the performance of selected DC/DA methods in four different test lengths. It was found that all methods had reasonably well estimated the indices, although the LL method had larger biases of PC and Kappa estimates in short tests, compared with the other three methods.

Study 2 focused on the test dimensionality and local item dependency (LID). Data

of various degrees of LID were generated. All methods greatly overestimated PA when the data had various levels of LID. The impact of LID on PC and Kappa estimates was much smaller, although the IRT-based methods tended to be more vulnerable in DC estimate to a high level of LID.

Study 3 checked the consequences of IRT model-data misfit on DC/DA estimates. Again PA was overestimated when the data were fitted with the incorrect model, while the PC and Kappa estimates received minimal impact from model misfit.

Study 4 checked the impact of using different scoring metrics. The scoring metric did not exhibit an obvious impact on DC/DA estimates, and the applicable methods performed in a similar way in the composite and theta score scales as in the raw score scale. Comparatively speaking, the LL method had a larger bias of PC and Kappa estimate on the composite score scale than on raw score scale. The HH method had consistently good estimates across the three scoring metrics.

5.2. Summary of the Findings

To summarize the findings in the simulation studies, it was found that

(1) The violation of model assumptions had a great negative impact on decision accuracy estimates, while had negligible impact on decision consistency estimates. Specifically speaking, when the data in the study had LID or model misfit, the "true" PA dropped noticeably but not "true" PC or Kappa. The "true" PA therefore became smaller than "true" PC, which was different from what was expected in the standard conditions. In addition, all selected methods had greatly over-estimated PA when data

had various degrees of LID, and slightly over-estimated PA when there was misfit between 3PL/GRM data and 1PL/PCM model. Since the conditions of local item independency and model fit are the fundamental assumptions of the models underlying the selected methods, violation of them would appear to be a threat to the validity of PA index.

While it was found that there were a couple of researches in literature looking at the factors affecting decision consistency, none of them studying the factors affecting decision accuracy. There were few papers investigating the PA index in simulation studies either. Since PA is important index indicating how accurate and valid the classification is, it is desirable that more studies would be conducted in the future to investigate the decision accuracy index and its related factors.

(2) Compared to the PA estimates, the PC and Kappa estimates had only minimal impacts from the above factors, probably because, whatever the problem with the data, it was consistent across the parallel forms of the test. Clearly test length was a bigger factor, but there was no differential impact across methods, although the LL method did not seem to perform as well at the other methods with short tests. Presumably the CTT assumptions were more consequential with short tests. Besides, it was found that IRT-based methods had poorer PC and Kappa estimates while LL_{strat} had the best performance when the data had a high level of LID.

(3) The results showed that scale for reporting was not important when the test was long, unidimensional, and had normal ability distribution.

The real data study was implemented as a supplementary approach to further investigate the performance of selected single-administration estimates of decision consistency and accuracy under different conditions. Combining the results of both studies, several conclusions can be drawn that reflect all of the work that was carried out in this research:

(1) The LEE and HH methods had almost identical results in both studies and across all conditions. This had been expected, as both approaches incorporate exactly the same assumptions. The LEE method provides an analytic solution, and the HH provides a simulation solution of the same approach. It was useful to see the closeness of the results. It is not so clear what might happen when sampling errors in the item parameter estimates are present.

(2) The LL method using standard Cronbach's alpha consistently under-estimated PC and Kappa indices in all conditions. The LL method using stratified alpha functioned noticeably better with higher reliability estimates and showed more robustness in short test length, LID and composite score. The studies indicated that the reliability estimate did have a great impact on the LL method, and a good estimate could be computed as long as an accurate reliability estimate was provided. The Cronbach's coefficient alpha, which is used the most widely in practice for the LL method, however, did not seem to be the best choice. Increases in the reliability estimates of even .05 in the real data study (due to the use of stratified alpha or parallel-form reliability) resulted in the LL method being notably more accurate.

(3) The LEE and HH methods had satisfactory performance and showed robustness of decision consistency estimates in most conditions. Furthermore, the HH method had a great flexibility and performed consistently well across different scoring metrics. One disadvantage of the HH method is that since it is simulation-based, every run would result in a different value for the estimate. However, the difference is small and should be negligible if the simulations were run multiple times and the average of the estimates is used. Besides, a large sample size is a must prior to any IRT application, including the HH method.

Lastly, it is worthy of pointing out that the IRT- and CTT-based methods make different assumptions about the parallel forms. IRT-based methods assume strictly parallel forms, where the items in parallel forms share exactly the same parameters, while CTT-based methods assume randomly parallel forms, where the items are randomly drawn from a parallel item bank. The "true" item parameters are fixed during the simulation studies, which may put an advantage for the IRT-based methods over the CTT-based methods. Therefore further simulation studies in which the "true" item parameters are varied by randomly drawing items from an item pool may be desired for future study.

BIBLIOGRAPHY

Algina, J., & Noe, M. (1978). A study of the accuracy of Subkoviak's single administration estimate of the coefficient of agreement using two true score estimates. *Journal of Educational Measurement, 15,* 101-110.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Berk, R. (1980). A consumer's guide to criterion-referenced test reliability. *Journal of Educational Measurement, 17,* 323-346.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley.

Bourque, M. L., Goodman, D., Hambleton, R. K., & Han, N. (2004). *Reliability estimates for the ABTE tests in elementary education, professional teaching knowledge, secondary mathematics and English/language arts* (Final Report). Leesburg, VA: Mid-Atlantic Psychometric Services.

Brennan, R. L. (2004). *BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy* (Version 1.0, CASMA Research Report No. 9). Iowa City: University of Iowa, Center for Advanced Studies in Measurement and Assessment.
Available at http://www.education.uiowa.edu/casma

Brennan, R. L., & Lee, W. (2006). *Correcting for bias in single-administration decision consistency indexes* (CASMA Research Report No. 18). Iowa City: University of Iowa, Center for Advanced Studies in Measurement and Assessment. Available at http://www.education.uiowa.edu/casma

Brennan, R. L., & Wan, L. (2004). *Bootstrap procedures for estimating decision consistency for single-administration complex assessments* (CASMA Research Report No. 7). Iowa City: University of Iowa, Center for Advanced Studies in Measurement and Assessment.Available at http://www.education.uiowa.edu/casma

Breyer, F. J., & Lewis, C. (1994). *Pass-fail reliability for tests with cut scores: A simplified method* (ETS Research Report No. 94-39). Princeton, NJ: Educational Testing Service.

Crocker & Algina (1986). *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth/Thomson Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L. J., Schoenemann, P., & McKie, D. (1965). Alpha coefficient for stratified-parallel tests. *Educational and Psychological Measurement, 25,* 291-312.

Ercikan, K., & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education*. *15*(3), 269-294.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education/Praeger.

Hambleton, R. K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W. R. Lenderking & D. Revicki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications* (pp. 57-78). Washington, DC: Degnon Associates.

Hambleton, R. K, & Novick, M. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, *10*(3), 159-170.

Hambleton, R.K, & Slater, S. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education*, *10*(1), 19-38.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Han, K. T., & Hambleton, R. K. (2007). User's Manual: WinGen (*Center for Educational Assessment Report No. 642).* Amherst, MA: University of Massachusetts, School of Education.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27*, 345-359.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*, 253-264.

Huynh, H. (1990). Computation and statistical inference for decision consistency indexes based on the Rasch model. *Journal of Educational Statistics, 15*, 353-368.

Kelley, T. L. (1947). *Fundamentals of statistics.* Cambridge, MA: Harvard University.

Lee, W. (2005). *Classification consistency under the compound multinomial model* (CASMA Research Report No. 13). Iowa City: University of Iowa, Center for Advanced Studies in Measurement and Assessment. Available at http://www.education.uiowa.edu/casma

Lee, W. (2007). Multinomial and compound multinomial error models for tests with complex item scoring. *Applied Psychological Measurement, 31,* 255-274.

Lee, W. (2008). *Classification consistency and accuracy for complex assessments using item response theory* (CASMA Research Report No. 27). Iowa City: University of Iowa, Center for Advanced Studies in Measurement and Assessment. Available at http://www.education.uiowa.edu/casma

Lee, W. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*(1), 1-17.

Lee, W., Brennan, R. L. & Wan, L. (2009). Classification consistency and accuracy for complex assessments under the compound multinomial model. *Applied Psychological Measurement, 33,* 374-390.

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*, 412-432.

Lee, W., & Kolen, M. J. (2008). *IRT-CLASS: A computer program for item response theory classification consistency and accuracy (Version 2.0)*. Iowa City, IA: University of Iowa, Center for Advanced Studies in Measurement and Assessment. Available at http://www.education.uiowa.edu/casma

Li, S. (2006). *Evaluating the consistency and accuracy of proficiency classifications using item response theory*. Unpublished dissertation. University of Massachusetts, Amherst, MA.

Liang, T., Han, K. T, & Hambleton, R. K. (2008). *ResidPlots-2: Computer software for IRT graphical residual analyses*, Version 2.0 [Computer Software]. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179-197.

Lord, F. M. (1965). A strong true score theory, with applications. *Psychometrika, 30,* 239-270.

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

No Child Left Behind Act of 2001. Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17,* 131-154.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating-scale data* [Computer program]. Chicago, IL: Scientific Software International, Inc.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5,* 343-355.

Peng, C. J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17,* 359-368.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel test. *Psychometrika, 30*, 39-56.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment Research & Evaluation*, 7(14). Available online: http://pareonline.net/getvn.asp?v=7&n=14.

Rudner, L. M. (2005). *Expected classification accuracy. Practical Assessment Research & Evaluation, 10*(13). Available online: http://pareonline.net/getvn.asp?v=10&n=13.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph Supplement, 17.*

Sireci, S. G, Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237–247.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13*, 265-276.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion referenced tests: a decision-theoretic formulation. *Journal of Educational Measurement*, *11*, 263-267.

Traub, R., & Rowley, G. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement, 4,* 517-545.

Wainer H, Wang X, Skorupski W. P., et al. (2005). A Bayesian method for evaluating passing scores: the PPoP curve. *Journal of Educational Measurement*, *2*(3), 271–281.

Wan, L. (2006). *Estimating classification consistency for single-administration complex assessments using non-IRT procedures*. Unpublished doctoral dissertation, The University of Iowa, Iowa City, IA.

Wan, L., Brennan, R. L., & Lee, W. (2007). *Estimating classification consistency for complex assessments* (CASMA Research Report No. 22). Iowa City: University of Iowa, Center for Advanced Studies in Measurement and Assessment. Available at http://www.education.uiowa.edu/casma

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement, 37*, 141-162.

Woodruff, D. J., & Sawyer, R. L. (1989). Estimating measures of pass-fail reliability from parallel half-tests. *Applied Psychological Measurement, 13,* 33-43.

Wu, M. (2004). Plausible values. *Rasch Measurement Transactions, 18*, 976-978.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing, 27,* 119-140.