1-1-2006

Evaluating the consistency and accuracy of proficiency classifications using item response theory.

Shuhong Li
University of Massachusetts Amherst

300 North Zeeb Road
P.O. Box 1346
Ann Arbor, Michigan 48106-1346 USA

800.521.0600 734.761.4700
web www.il.proquest.com

Printed in 2006 by digital xerographic process
on acid-free paper

EVALUATING THE CONSISTENCY AND ACCURACY OF PROFICIENCY CLASSIFICATIONS USING ITEM RESPONSE THEORY

A Dissertation Presented

by

SHUHONG LI

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

DOCTOR OF EDUCATION

May 2006

School of Education

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

EVALUATING THE CONSISTENCY AND ACCURACY OF PROFICIENCY CLASSIFICATIONS USING ITEM RESPONSE THEORY

A Dissertation Presented

by

SHUHONG LI

Approved as to style and content by:

____
Stephen G. Sireci, Chair

____
Ronald K. Hambleton, Member

____
Aline G. Sayer, Member

____
Christine B. McCormick, Dean
School of Education

EVALUATING THE CONSISTENCY AND ACCURACY OF PROFICIENCY CLASSIFICATIONS USING ITEM RESPONSE THEORY

A Dissertation Presented

by

SHUHONG LI

Approved as to style and content by:

Stephen G. Sireci, Chair

Ronald K. Hambleton

Ronald K. Hambleton, Member

Aline G. Sayer

Aline G. Sayer, Member

Christine B

Christine B. McCormick, Dean
School of Education

ACKNOWLEDGEMENTS

I would like to extend my heartfelt thanks to the following people who have assisted me in the odyssey of writing my dissertation, as well as throughout my years in the Research and Evaluation Method Program at the University of Massachusetts Amherst.

First, I was very fortunate to have Steve Sireci, Ronald Hambleton and Aline Sayer as my committee members. Steve, your support, insight, comments, and suggestions for the multiple versions of this project were invaluable. Ron, I cannot thank you enough for your perspective and guidance that were crucial in the shaping of this dissertation. Aline, I am grateful to you for all your great advice and expertise, especially for your patience, understanding, and willingness to accommodate that helped ease this whole process of writing my dissertation.

I am forever indebted to the faculty members of REMP for the education I received from them. All of them either read the whole or part of my dissertation, or talked with me about my dissertation. Among many other things, I thank Ron for admitting me into this outstanding program, for his always prompt email replies, for his admirable work ethic, and most importantly, for the precious opportunities of working with him on several projects that were instrumental in my academic development. I had the privilege of having Swami as my initial advisor for only a year, but his influence will be a constant source of inspiration to me. Most specially, I am deeply indebted to my advisor Steve, for all his patient and generous guidance in academics as well as in communicating and cooperating with colleagues, and for all the opportunities he provided me (including my role of being Godzilla's translator in the NCME satirical session). I thank Lisa Keller and Jane Rogers for their inspiring lectures and all the care and concern they showed me. Last, but not least, I'd like to gratefully acknowledge Craig Wells for his most generous mentoring in my research activities, and for his great and kindest efforts in boosting my confidence whenever I doubted myself. I am grateful too for all the assistance I got from Peg throughout my years at REMP.

I would also like to thank all my fellow REMP classmates and friends in Amherst. Especially to Ning Han, I am indebted to you for all your tremendous help, and the encouragement, support, and hospitality (including the countless meals) I have enjoyed from your family. Together with Xiaoying's family, you made my life in Amherst so much easier throughout these years. To my witty, considerate, and intelligent officemates Jeff and Peter, I enjoyed your friendship, companionship, jokes, and all the fights we had; you will always be part of my fondest memories about my student life. I must extend my gratitude and deep appreciation to Stephen for his constant friendship and for all his invaluable help and support. I should also thank my dear friend Mingxuan and for all her care and precious friendship.

To my beloved family members, I thank all of you for your unconditional love and support that sustained me and motivated me to press forward even at the hardest times for all these years, though you are half a world away. I could not have done this without you; therefore, I dedicate this degree to all of you.

ABSTRACT

EVALUATING THE CONSISTENCY AND ACCURACY OF CLASSIFICATIONS IN ITEM RESPONSE THEORY

MAY 2006

SHUHONG LI, B.A., XINJIANG NORMAL UNIVERSITY
M.A., XI'AN INTERNATIONAL STUDIES UNIVERSITY
Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST
Directed by: Professor Stephen G. Sireci

As demanded by the No Child Left Behind (NCLB) legislation, state-mandated testing has increased dramatically, and almost all of these tests report examinee's performance in terms of several ordered proficiency categories. Like licensure exams, these assessments often have high-stakes consequences, such as graduation requirements and school accountability. It goes without saying that we want these tests to be of high quality, and the quality of these test instruments can be assessed, in part, through the decision accuracy (DA) and decision consistency (DC) indices.

With the popularization of IRT, an increasing number of tests are adopting IRT for test development, test score equating and all other data analyses, which naturally calls for approaches to evaluating DA and DC in the framework of IRT. However, it is still common to observe the practice of carrying out all data analyses in IRT while reporting DA and DC indices derived in the framework of CTT. This situation testifies to the necessity to the exploration of possibilities to quantify DA and DC under IRT.

The current project addressed several possible methods for estimating DA and DC in the framework of IRT with the specific focus on tests involving both dichotomous and polytomous items. It consisted of several simulation studies in which the all IRT methods introduced were valuated with simulated data, and all methods introduced were also be applied in a real data context to demonstrate their application in practice. Overall, the results from this study provided evidence that would support the use of the 3 IRT methods introduced in this project in estimating DA and DC indices in most of the simulated situations, and in most of the cases the 3 IRT methods produced results that were close to the "true" DA and DC values, and consistent results to (sometimes even better results than) those from the commonly used L&L method. It seems the IRT methods showed more robustness on the distribution shapes than on the test length. Their implications to educational measurement and some directions for future studies in this area were also discussed.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

As demanded by the No Child Left Behind (NCLB) legislation (2002), state-mandated testing has increased dramatically, and almost all of these tests report examinee's performance in terms of several ordered proficiency categories. These proficiency categories are often determined by a standard setting process, which can provide easy interpretations and clear verbal descriptions of student performance. The same is true to some other national and commercially marketed standardized achievement tests such as the National Assessment of Educational Progress (NAEP), and the Terra Nova-CTBS series (CTB/McGraw-Hill, 1996).

Like licensure exams, these assessments often have high-stakes consequences, such as graduation requirements and school accountability, which can have significant effects on students, teachers, schools, and administrators. Moreover, there is federal legislation linking funding to standardized test score improvement at grades 3 through 8, and this has the potential to increase the stakes associated with standardized tests throughout the nation. It is clear that we want these tests to be of high quality. The quality of these test instruments can be assessed through the decision accuracy (DA) and decision consistency (DC) indices, the origin of which can be traced to the research on the criterion-referenced tests.

The importance of classification accuracy and consistency has long been widely recognized. For instance, the standard 2.15 in the current Standards for Educational and Psychological Testing (American Educational Research Association, American

Psychological Association, & National Council on Measurement in Education, 1999) states 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 …” A number of studies have been devoted to the demonstration of the procedures for quantifying the indices of DA and DC.

As noted by Hambleton and Slater (1997), before 1973, it was common to report a KR-20 or a corrected split-half reliability estimate to support the use of a credentialing examination. Since these two indices only provide estimates of the internal consistency of examination scores, Hambleton and Novick (1973) suggested that the DC could be defined in terms of the consistency of candidate decisions resulting from two administrations of the same examination or parallel forms of the examination. Swaminathan, Hambleton, and Algina (1974) followed up with a more detailed explanation of DC that utilized Cohen's kappa. As compared with the definition of DC, DA is the extent to which the actual classifications of the test takers agree with those that would be made on the basis of their true score, if their true scores could somehow be known (Livingston & Lewis, 1995).

The definition by Hambleton and Novick (1973) pointed to a new direction to evaluating the DC index: an index of reliability which reflects the consistency of classifications across repeated testing. Since then, several methods have been put forward to calculate such an index. For example, a raw index of reliability represented by the proportion of consistent mastery/mastery and nonmastery/ nonmastery classifications on two tests with cutting score c, symbolized p_c, was proposed by

Swaminathan, Hambleton, and Algina, (1974, 1975). Since data for such repeated testings are seldom available in practice, several researchers proposed procedures for estimating classification consistency indices using test scores obtained from a single test administration by imposing psychometric models on test scores such as Huynh (1976), Subkoviak (1976), and Hanson and Brennan (1990). Since the previous methods all deal with binary data, Livingston and Lewis (1995) came up with a method that can be used with data including either dichotomous data, polytomous data, or the combinations of two.

All the previously described methods are calculated in the framework of classical test theory (CTT). CTT has been the mainstay of psychological test development for most of the 20th century, and it has defined the standard for test development, beginning with the initial explosion of testing in the 1930s. However, since Lord and Novick's (1968) classic book introduced model-based measurement, a quiet revolution has occurred in test theory. Item response theory (IRT) has rapidly become mainstream for psychological measurement. Increasingly, standardized tests are developed from IRT due to the more theoretically justifiable measurement principles and the greater potential to solve practical measurement problems (Embertson & Reise, 2000). With the popularization of IRT, the evaluation of DA and DC under IRT began to attract the interest of researchers. For example, Rudner (2001, 2004) introduced his method for evaluating DA in the framework of IRT.

1.2 Statement of the Problem and Its Significance

IRT is a powerful, model based scaling technique that has several useful applications in educational and psychological testing contexts (Hambleton, Swaminathan, & Rogers, 1991), and the theoretical benefits of IRT have had a tremendous impact on various practical aspects of the measurement process. An increasing number of test developers are employing IRT for test development, test score equating and all other data analyses, but there is a lack of research and application of methods for calculating DA and DC in the context of IRT. For example, all of methods mentioned previously (except for the Rudner method) are carried out in the framework of CTT. It is still common to observe the practice of carrying all data analyses in IRT while reporting DA and DC indices derived in the framework of CTT, and this situation testifies to the necessity to further the exploration of more possibilities to qualify DA and DC under IRT. In addition, all CTT methods have appeared to be fairly complex and labor intensive to calculate. Given the availability of IRT software, it would be interesting to see if an easier and more convenient solution could be found.

1.3 Purpose of the Study

The purpose of the study is to review previous methods and introduce new methods for addressing DA and DC in the framework of IRT with the specific focus on tests involving both dichotomous and polytomous items. Specifically, answers to the following 3 questions were investigated: (1) How can the Rudner method be extended to include the calculation of DC in addition to DA? (2) How can IRT methods provide DA and DC results on the raw test score metric? How well does the Hambleton and Han (2005) IRT method function on the raw test score metric? and (3) How do these IRT methods compare to the most widely used classical approach using both simulated and real data?

1.4 Outline of the Study

The following is a description of a study investigating possible methods for estimating DA and DC in IRT. The description begins with a comprehensive review of previous research into other methods of estimating DA and DC in both CTT and IRT, which is warranted for illustrating why the methods under investigation in the present project are important and how they differ from previous methods. Subsequent to the review of literature is a detailed description of the method of the study. This includes three studies that detail explicitly the description of the three sets of IRT methods, delineate the corresponding procedures for applying them, and evaluate performance of the 3 IRT methods using a simulation study. This paper concludes with the presentation of the results, their implications identification to educational measurement and some directions for future studies in this area.

CHAPTER 2

LITERATURE REVIEW

This chapter presents a comprehensive review of previous research into all existing methods of estimating DA and DC in both CTT and IRT. It begins with the introduction of classical approaches to reliability estimation and the elaboration of why these approaches are not appropriate to for use with criterion-referenced testing (CRT). It then describes the origin of the concepts of DA and DC, which is followed by the delineating of existing approaches to DA and DC in both CTT and IRT. At the end, the chapter summarizes the major findings gleaned from the review of the literature, and explains how the findings link to the research questions put forward in Chapter One.

2.1 Classical Approaches to Reliability Estimation

2.1.1 CTT Methods

It has been more than 90 years since Charles Spearman invented the concept of reliability. Reliability of test scores refers to the consistency of test scores over time, over parallel forms, or over items within the test. It follows naturally from this definition that calculation of reliability indices would require a single group of examinees taking two forms of a test or even a single test a second time, but this is often not realistic in practice. Thus, it is routine to report single-administration reliability estimates such as corrected split-half reliability estimates and/or coefficient alpha. Accuracy of test scores is another important concern that is often checked by comparing test scores against a criterion score, and this constitutes a main aspect of validity (Hambleton & Li, 2005).

As suggested by its name, in the split-half method, one form of a test is administered to a group of examinees, the items in the test are grouped into two subtests to create two half-tests that are as close to parallel as possible, and then the two half-tests are scored separately for each examinee. Crocker and Algina (1986) documented the following four methods that can be employed to divide the test as such:

1) Assign all odd-numbered items to form 1 and all even-numbered items to form 2.
2) Rank order the items in terms of their difficulty levels (p-values) based on the responses of the examinees; then assign items with odd-numbered ranks to form 1 and those with even-numbered ranks to form 2.
3) Randomly assign items to the two half-test forms.
4) Assign items to half-test forms so that the forms are "matched" in content.

Ideally, if the two half-tests are strictly parallel, the correlation between the scores represents an estimate of the reliability of either half-test. Since what is desired is the reliability of scores on the whole test, the coefficient of correlation thus derived is corrected by using the Spearman-Brown formula to derive the reliability coefficient had two whole tests administered. Rulon (1936) also proposed a method for deriving split-half reliability without having to use the Spearman-Brown correction.

The split-half procedure is known to have a drawback: it does not yield a unique estimate of the test's reliability coefficient, because there are many possible ways to divide a test into halves (Crocker & Algina, 1986). To overcome this limitation, the procedures such as coefficient alpha for polytomous response data, and the KR-20 and KR-21 formulas for binary data were proposed and have been very widely used till today. It can be seen that both split-half and coefficient alpha make use of the information contained in the item scores to yield an index of the internal consistency of the examinees' responses to the items within a single test form.

2.1.2 Limitations of Using CTT Reliability Estimates

Glaser (1963) held that scores on achievement tests can be interpreted in two ways. One way is in terms of the relative position of an examinee's score in comparison with a group of examinees who have also taken the test, for example, the reporting of test performance in the form of percentiles or as z-scores. The second way is in terms of the degree to which the student has attained the goals of instruction. This reveals the necessity of differentiating two kinds of tests: norm-referenced testing (NRT) and criterion-referenced testing (CRT), which serve different purposes and have different implications on test development and test evaluation.

Norm-referenced testing is designed mainly to distinguish or compare examinees on the construct measured by the test. With these kinds of tests, test score reliability is judged by the stability of the examinee rankings or scores over parallel-form administrations or test-retest administrations of the test, and the specific procedures designed for doing this included the split-half and coefficient alpha methods. Examinees are basically rank-ordered based on their test scores. For the rank ordering to be reliable, the test itself needs to spread out the examinees so that the present measurement errors do not distort the ranking that would be obtained if true scores had been used. Items of middle difficulty and high discriminating power are usually preferred in order to spread out the examinee scores or to maximize the test score variability, given constraints on such things as test content and test length (Hambleton & Li, 2005).

Criterion-referenced testing, however, was introduced in the United States in the 1960s as a response to the need for assessments that could determine what persons knew and could do in relation to a well-defined domain of knowledge and skills, rather than in relation to other persons. With the CRT score information, the level of proficiency of candidates can be determined, and in many cases, diagnostic information can be provided that will be helpful to candidates in working on their weaknesses (Hambleton & Li, 2005).

Today, the uses of CRTs are wide-spread in education, the military, and industry. With criterion-referenced credentialing exams, normally only two performance levels are used: passing and failing; with many state criterion-referenced tests, examinees, based upon their test scores, are assigned to one of four performance levels: Failing, Basic, Proficient, and Advanced (Hambleton & Li, 2005). Performance levels are obtained by applying two or several performance standards on the reporting scale, and these performance standards are often derived from the process of standard setting, which are some points on the reporting scale that are used to sort examinees into the desired performance levels.

Unlike norm-referenced testing, in criterion-referenced testing we do not necessarily desire highly variable test scores. Therefore, internal consistency reliability may not be informative. The decision of whether or not a student's performance is satisfactory so that it can be regarded as passing the test should be made independent of other students' performance. Despite the differences between the NRTs and the CRTs, when CRTs were first introduced, test score reliability for this kind of testing was still calculated with the split-half or coefficient alpha procedures, which are the techniques developed largely for NRTs. Given the differences lying in the nature of the two tests, some scholars realized the necessity for the CRTs to have their own techniques in evaluating reliability indices. Hambleton and Novick (1973) first introduced the concept of DC. They argued that indices providing estimates of the internal consistency of examination scores should not be used with the CRTs. Since this new aspect of evaluating DC for the CRTs was proposed, the literature on the subject of CRT reliability has expanded year by year in apparent conformity with an exponential function (Traub, 1994). This definition conveys the message that in evaluating the quality of the CRTs, it is the quality of the decisions based on test scores that is more of concern, not the reliability of test scores as it is traditionally understood.

2.2 The Concepts of DA and DC

As has been described, with CRTs examinee performance is typically reported in performance categories, so reliability and the validity of the examinee classifications is of greater importance than the reliability and validity associated with test scores. That is, the consistency and accuracy of the decisions based on the test scores should outweigh the consistency and the accuracy of test scores with CRTs.

Being cognizant of the above critical point, Hambleton and Novick (1973) suggested that the classification consistency could be conceptualized as consistency in the making of pass and fail decisions, and can therefore be defined in terms of the consistency of candidate decisions resulting from two administrations of the same examination or parallel forms of the examination, that is, an index of reliability which reflects the consistency of classifications across repeated testing. According to this definition, a good test is one that classifies examinees in the same way if the examinees can take a repeated or a parallel form of the test. Reliability or DC index in this situation can be high even though the score variability across all test takers is small.

Traub (1994) pointed out that this notion of DC advanced by Hambleton and Novick (1973) underlies a new conception of reliability for criterion-referenced tests, and was operationalized by Swaminathan, Hambleton, and Algina (1974). The publication of the Hambleton and Novick (1973) paper, together with one by Livingston (1972), spawned a large body of literature on the formulation of new indices of reliability for CRTs.

As compared with this definition of DC, DA refers to the "extent to which the actual classifications of the test takers agree with those that would be made on the basis of their true score, if their true scores could somehow be known" (Livingston & Lewis, 1995). If we say DC reflects the reliability of proficiency classification designations, DA then refers to the validity of proficiency classifications. Today they both occupy a very important role in the standard-based performance.

It is worth pointing out that the value of DA is higher that of DC. This is because the calculation of DA is based on one set of observed scores and one set of true scores which are supposed to be free of any measurement error due to improper sampling test questions, flawed test items, problems with the test administration and so on. While in calculating DC, two sets of observed scores are involved. The factors that influence decision-making include the method of assigning examinees to mastery states, selection of the cutting score, test length, and heterogeneity of the group.

The levels of DC and DA required in practice will depend on the intended uses of the CRT and the number of performance categories. There have not been any established rules or guidelines to help determine the levels of DC and DA needed for different kinds of educational and psychological assessments. In general, the more important the educational decision to be made, the higher the consistency and accuracy need to be (Hambleton & Li, 2005).

2.3 Methods of Estimating DC and DA

As has been said, the measurement literature includes several methods for calculating DA and DC. There are methods devoted to dichotomous items only, methods to both dichotomous and polytomous, methods for binary decisions, and methods for multiple decisions. All of these methods (except for the Rudner method) are carried out in the framework of classical test theory (CTT). A complete set of approaches for estimating DA and DC are contained in Table 2.3.1, which is followed with a brief review of each of the method tabulated.

2.3.1 The Hambleton and Novick Procedure

The new concepts of DA and DC by Hambleton and Novick (1973) can be visually displayed in Figure 2.3.1.1 and Figure 2.3.1.2. Their procedures are obviously for double administrations and the tests include dichotomous items only.

In Figure 2.3.1.1, the classifications are consistent when the decisions made based on Test One agree with those made based Test Two, which are represented by the diagonal cells. Inconsistent classifications occur when the classifications based on the two tests are not in agreement, i.e., a student is classified as a Master on one test form, but is classified into the Non-Master category on another test form. The inconsistent classifications are denoted by the off-diagonal cells. In Figure 2.3.1.2, consistent classifications happen when the decisions made based on the true scores are in accordance with those made based on the observed scores. To be specific, the diagonal cells again represent consistent decisions and the off diagonal cells represent those inconsistent decisions.

2.3.2 Swaminathan, Hambleton and Algina Procedure

Swaminathan, Hambleton and Algina (1974) extended the Hambleton-Novick to the case where there were k not just two performance categories, and their DC index is presented by the following formula:

$$p_0 = \sum_{i=1}^{k} p_{ii} \tag{1}$$

where p_{ii} is the proportion of examinees consistently assigned to the ith performance categories (or levels) across the two administrations. It can be seen that the upper bound of this agreement coefficient is 1.00, which occurs if classifications made on both administrations are consistent for all examinees in the group, and the lower bound of this agreement coefficient represents the proportion of consistent classifications made by chance if decisions made on the second administration were completely independent of decisions made on the first administration.

Swaminathan, Hambleton and Algina (1974) further argued that reliability should be defined as a measure of agreement over and above that which can be expected by chance between the decisions made about examinee mastery states in repeated test administrations for each objective measured by the criterion-referenced test. So in order to correct for chance agreement, they created a statistic based on Cohen's kappa (1960) which is a generalized proportion agreement index frequently used to estimate inter-judge agreement::

$$k = \frac{p - p_c}{1 - p_c} \tag{2}$$

where p is the proportion of examinees classified in the same category by two administrations, and p_c is the agreement by chance. The upper bound of kappa is 1.00, which occurs when the decisions made based on the two test administrations completely agree with each, and the lower bound of kappa is 0, which happens when the decisions made based on the two administrations are completely independent of each other.

This paper is important because it not only realized the new concept of DA and DC, but also initialized the practice of reporting DC that has been adopted till today: the reporting of both an agreement coefficient and the kappa coefficient which is the agreement coefficient corrected by chance. Berk (1980) noted that Hambleton and Novick's (1973) two-administration method for estimating DC is the easiest to understand, to compute, and to interpret, and it appears to have the greatest utility for classroom test construction and decision making. Swaminathan et al.'s (1974) two-administration method for estimating kappa also merits the attention of test publishers and test makers at the district and state levels. The disadvantage of the two methods is that they both require two test administrations.

2.3.3 Subkoviak's Procedure

The concepts of DC and kappa were quickly accepted by the measurement field for use with CRTs. However, the previously introduced methods such as kappa are defined by means of a test retest method with the same test form or a parallel one. Since repeated testing is often an impractical restriction, it would be desirable to have ways to estimate the reliability of decisions on the basis of a single testing. Subkoviak's procedure (1976) was designed to serve this purpose, and is suitable for tests composed of dichotomous items only.

Subkoviak's method (1976) is based the statistical theory for binomial variables, i.e., it is assumed that observed scores are independent and distributed binomially with two parameters: the number of items and the examinee's proportion-correct true score. In his procedure, the examinee's proportion-correct true score can either be estimated by the quantity maximum likelihood estimation or linear regression. The procedure estimates the true score for each individual examinee at a time without making any distributional assumptions for true scores. When combined with the binomial or compound binomial error model, the estimated true score will provide a consistency index for each examinee, and averaging this index over all examinees gives the DC index.

Subkoviak's procedures have an assumption that each examinee has a constant proportion correct value for all items in the estimation of DC. Spray and Welch (1990) noted that no systematic research had been carried out on the effect of the violation of this assumption. They conducted a study to examine the effect that large within-examinee item difficulty variability has on estimates of the proportion of consistent classification of examinees with Subkoviak procedure. Analyses of simulated data revealed that the use of a single estimate for an examinee's probability of a correct response, even when that probability varied greatly within a test for an examinee, did not affect the estimation of the proportion of consistent classifications. They pointed out that because the Subkoviak and Huynh (1976) procedures produce similar results, one would assume that this would be the case using the beta-binomial density model as well.

2.3.4 Huynh Procedure

Huynh's (1976) method represents another single-administration estimator of DC. His procedure is based on a more complicated theory than Subkoviak's, which is known to be his two-parameter "bivariate beta-binomial model". This model relies on the assumption that a group of examinees' ability scores follow the beta distribution with parameters α and β, and the frequency of the observed test scores x follow the beta-binomial (or negative hypergeometric) distribution with parameters α and β. The model is defined by the following:

$$f(x) = \frac{n!}{x!(n-x)!} \mathrm{B}(\alpha + x, \alpha + \beta - x) / \mathrm{B}(\alpha, \beta) \tag{3}$$

where n is the total number of items in the test, and B is the beta function with parameters α and β, which can be estimated either with the moment method – making use of the first two moments of the observed test scores – or with the maximum likelihood (ML) method described in his paper. The probability that an examinee has been consistently classified into a particular category can then be calculated by using the beta-binomial density function. Hanson and Brennan (1995) extended Huynh's approach by using the four-parameter beta distribution for true scores.

Berk (1980) pointed out that Huynh's (1976) beta-binomial model has several distinct advantages compared to the alternatives. It is based on the mathematically elegant Keats and Lord model (1968). The DC index resulting from this model is easily interpretable, and the violation of equal item difficulty seems to have negligible effect on estimates (Subkoviak, 1978). The disadvantage of the approach is that it is one of the most conceptually, mathematically, and computationally complex approaches. Peng and

Subkoviak (1980) provided a simple normal approximation of the beta-binomial distributions that can be hand calculated. They found this approximation to provide relatively accurate estimates of agreement coefficient and kappa coefficient. They justified this approximation by noting that there is support in the literature for the possibility of approximating the beta-binomial family of distributions with the normal family. According to them, the beta-binomial family can be approximated by the negative binomial family, which, in turn, could be approximated by the normal family.

2.3.5 Livingston and Lewis Procedure

DC can be indexed relatively easily in the situation where we only need to deal with binary data. However, the increasing popular "Mixed format" tests (i.e., tests including dichotomous and polytomous items) call for DC and DA calculation procedures that can be used with this kind of test.

Livingston and Lewis' (1995) procedure can be used with data that is either dichotomous, polytomous, or the combination of the two. It involves estimating the distribution of the proportional true scores *Tp* using the strong true score theory. This theory assumes that the proportional true score distribution has the form of a four parameter beta distribution with density

$$g(T_p / \alpha, \beta, a, b) = \frac{1}{Beta(\alpha + 1, \beta + 1)} \frac{(T_p - a)^{\alpha} (b - T_p)^{\beta}}{(b - a)^{\alpha + \beta + 1}} \qquad (4)$$

where *Beta* is the beta function, and the four parameters of the function can be estimated by using the first four moments of the observed scores for the group of examinees. Then the conditional distribution of scores on an alternate form (given true score) is estimated using a binomial distribution.

2.3.6 Emerging IRT Procedures

All of the previous described methods operate in the framework of CTT. From 1980 onwards, IRT has been increasingly used in test development and all other data analyses such as item analysis and test score equating. The necessity and attractiveness of IRT methods for evaluating DC and DA are being realized, though the literature reveals that there is still a lack of research in this area. For example, Rudner (2001, 2004) introduced his method for evaluating DA in the framework of IRT, and Hambleton and Han (in Bourque, et. al., 2004), and these two procedures will be described in detail in the following two sections.

2.3.6.1 Rudner Method

Rudner (2001) proposed a procedure for computing expected classification accuracy for tests consisting of dichotomous items and later extended the method to tests including polytomous items (see, Rudner, 2004). It should be noted that Rudner referred to θ and $\hat{\theta}$ as 'true score' and 'observed score', respectively in his papers. He pointed out that for any given true score θ, the corresponding observed score $\hat{\theta}$ is expected to be normally distributed, with a mean θ and a standard deviation of $se(\hat{\theta})$. The probability of an examinee with a given true score θ of having an observed score in the interval $[a,b]$ (an interval between two cut scores) on the theta scale is then given by

$$p(a<\hat{\theta}<b\,|\,\theta)=\phi\left[\frac{b-\theta}{se(\hat{\theta})}\right]-\phi\left[\frac{a-\theta}{se(\hat{\theta})}\right] \quad , \qquad (5)$$

where $\phi(Z)$ is the cumulative normal distribution function. He noted further that multiplying equation (1) by the expected proportion of examinees whose true score is θ yields the expected proportion of examinees whose true score is expected to be in interval $[a,b]$, and summing or integrating over all examinees in interval $[c,d]$ (an interval between two cut scores) gives us the expected proportion of all examinees that have a true score in $[c,d]$ and an observed score in $[a,b]$. If we are willing to make the assumption that the examinees' true scores (θ) are normally distributed, the expected proportions of all examinees that have a true score in the interval $[c,d]$ and an observed score in the interval $[a,b]$ are given by

$$\sum_{\theta=c}^{d} P(a<\hat{\theta}<b|\theta)f(\theta)=\sum_{\theta=c}^{d}\left[\phi\left[\frac{b-\theta}{se(\hat{\theta})}\right]-\phi\left[\frac{a-\theta}{se(\hat{\theta})}\right]\right]\Phi(\frac{\theta-\mu_{\theta}}{\delta_{\theta}}), \qquad (6)$$

where $se(\theta)$ is the reciprocal of the square root of the Test Information Function at theta, which is the sum of the Item Information Function, and $f(\theta)$ is the standard normal density function $\Phi(Z)$ (Rudner, 2004).

Suppose we have a test that is measuring a single latent ability θ, and students are divided into k categories based on the test. Also, we have calibrated the test by adopting appropriate IRT model(s) and have gotten the students' ability estimate $\hat{\theta}$ (or observed scores in Rudner's terminology). Let c_0 denote the minimum possible score for the test, c_k the maximum possible score for the test, and c_i $(i \neq 0$ or $k)$ the ith cut-off score. Table 2.3.6.1.1 constitutes the contingency table for calculating DA indices with Rudner's method.

Since this is the first IRT method introduced, the procedures to operationalize the method are also provided as follows:

1) If the cut-off scores are set on the raw score scale, as they usually are, the first step

will be to transform the cut-off scores from the raw score scale to the theta score scale. This step can be implemented by making use of the Test Characteristics Curve (TCC); that is, each cut-off score can be mapped to a corresponding θ score via the TCC of the test.

2) From step 1), we have $\theta_{c_0}, \theta_{c_1}, \theta_{c_{i-1}}, \theta_{c_i}, \ldots \theta_{c_{k-1}}$, which classify the students into k categories based not only on the observed theta scores $(\hat{\theta})$ but also on students' true scores which are assumed to be normally distributed.

3) A $k \times k$ contingency table for classification table is set up by either having the categories for the observed scores as the rows and the categories for the true scores as the columns (as in this paper) or vice versa. For example, in this table, p_{11} refers to the proportion of the students who are classified into the first category based on both observed scores and their true scores.

4) Next, the elements of the contingency table that are conditional probabilities are calculated as follows:

$$p_{ij} = \int_{\theta_{c\,j-1}}^{\theta_{c\,j}} p(\theta_{c_{i-1}} \leq \hat{\theta} < \theta_{c_i}) f(\theta) d\theta \text{, where } 0<i<k, 0<j<k. \quad (7)$$

When $i = j$, $$p_{ii} = \int_{\theta_{c_{i-1}}}^{\theta_{c_i}} p(\theta_{c_{i-1}} \leq \hat{\theta} < \theta_{c_i}) f(\theta) d\theta \,. \quad (8)$$

(4) For example, when $i = j = 1$,

$$p_{11} = \int_{\theta_{c0}}^{\theta_{c1}} p(\theta_{c_0} \leq \hat{\theta} < \theta_{c_1}) f(\theta) d\theta \,. \quad (9)$$

5) The overall DA index is calculated as the sum of the diagonal elements in the contingency table.

2.3.6.2 Hambleton and Han Method

Bourque and et. al. (2004) documented another IRT method for evaluating DA and DC by Hambleton and Han for 0-1 data. According to the authors, Hambleton and Han were provided with only the following for them to carry out the task of calculating the DA and DC for a particular test instrument: (1) item parameter estimates (which in their case were b-values because Rasch model was adopted for the calibration of the test) for each form of each test, (2) ability scores for candidates, (3) two cut scores on the ability score metric for each test, and (4) CTT reliability estimate of the test.

In their situation, the best known single administration estimates of DA and DC we have in the measurement literature (for example, Subkoviak, 1976; Huynh, 1976; Livingston and Lewis, 1995) are all implemented on the test score metric and therefore could not be adopted. They had to come up with a three-step IRT-based method for indexing single administration DA and DC for tests including dichotomous items with binary or multiple decisions. The method is described by the authors as follows:

(1) Estimation of the ability score distribution and generation of the observed scores. As demanded by the definition of DA and DC, they needed an estimate of the "ability score distribution" from which examinees can be chosen and examinee performance on parallel-forms of a test can be simulated. As has been mentioned, the two authors were provided with ability estimates for the examinees, but according to them, this distribution was somewhat more heterogeneous than the distribution of interest. They noted that this is analogous to having observed score distributions when

true score distributions are needed when working within a classical measurement framework It should be pointed out that in practice, users would often not be provided with ability estimates. Rather, they would start with a response matrix and proceed to estimate item parameter and ability estimates by applying appropriate IRT models.

In order to better approximate the ability score distribution, Kelley regressed estimates of ability were substituted for ability score estimates to reduce ability score variability. This process was depicted by the following formula:

$$\theta_i = \hat{\theta}_i * \rho + (1-\rho) * \mu_{\hat{\theta}_i} \tag{10}$$

where $\hat{\theta}$ is an observed theta estimate, ρ is the CTT reliability estimate, and $\mu_{\hat{\theta}_i}$ is the mean for the observed theta estimates. Hambleton and Han explained that by applying formulae (10), the ability score estimates were pulled in a bit toward the mean of the ability score distribution. Also, the amount of "regression" will be limited if the reliability estimate is high.

For each regressed estimate of ability, and with knowledge of the item parameter estimates for one of the forms, they generated both item scores and a total test score (X) for each candidate. The way to generate item scores (or response patterns) is a routine process using Monte-Carlo procedures that requires item statistics (for example, item b-values if Rasch model is used), and candidate ability scores. In this process, a probability *p* is to compare with a [0, 1] uniform distribution random number *r*. If $p > r$, the event with the probability *p* happens. Otherwise the event does not happen. With dichotomous items, an event that happened means the examinee scores 1 on the item, and an event that did not happen means the examinee scores 0 on the item. With item scores in hand for a candidate, they can be summed up to obtain a total test score. Now it would be an easy task to produce a second test score (Y), again using the regressed estimate of ability for each candidate and the item statistics (i.e., b-values) for the items in the test form. At this point, for each candidate who took the form of the test they were simulating, they had (model-based) test scores on parallel-forms, denoted test score X and test score Y.

(2) Transforming and/or applying cut scores. The authors were provided with two cut scores for the test on the ability scale, so these cut scores were transformed to the test score scale by making use of the test characteristic curve which can be produced from the available test form item statistics. By comparing these cut scores on the test score scale, candidates could be assigned to performance categories using test X and test Y scores. By mapping over the regressed estimates of ability to the test score scale, an estimated true score for each candidate could be obtained also.

(3) Calculation of DC, kappa, and DA indices. At this point in the process, for each candidate on a form of the test, the following things have been provided: (1) ability score estimate, (2) regressed ability estimate, (3) raw score on test X, (4) performance classification using test score X, (5) raw score on test Y, (6) performance classification using test score Y, (7) estimated true score, and (8) true classification of the candidate (obtained using 7, and the cut scores on the test score scale). With this information in hand, it was straightforward to calculate the consistency of performance classifications or DC (using 4 and 6), kappa (using 4 and 6), and accuracy of performance classifications (using 4 and 8, and also 6 and 8).

The authors noted some advantages of their approach to produce DA and DC estimates. First, it is considerably easier to write code for this approach compared to the classical test theory methods such as the Livingston and Lewis method, which effectively relies on an analytic solution to obtain estimates but requires some very complicated modeling of the data and strong assumptions about score distributions. Second, no assumptions about the distributions of test scores and ability scores needed to be made, except for the assumption that the IRT model employed adequately fits the data. They further noted that in practical work it is common to demonstrate that the estimated test scores, assuming the model to be true, approximate the observed test scores. Therefore, in their work they assumed that the model fit the data, and so the estimated scores for tests X and Y were taken to be the actual scores. The disadvantage of this method is that it is based on simulation; therefore, the DA, DC and Kappa values derived from each calculation might vary a little bit. They suggested that the calculations be performed multiple times and the mean values of the DA, DC and Kappa indices be reported as the final ones.

2.4 Topics Covered in the Literature

The emergence of the concepts DC and DA and the subsequent CTT methods for evaluating these indices triggered many studies in this area, most of which appeared in late 1970's and early 1980's. Roughly, these studies centered on three topics: review of the methods proposed, certain follow up studies to evaluate the accuracy of the methods proposed, and comparison of the methods.

The papers on reviewing the literature that have been located all appeared in the 1980's, and all address the methods in CTT. The authors would typically go over the methods that had been put forward, and summarize their respective advantages and disadvantages gleaned from the literature. Good examples of this type of review are Traub and Rowley (1980) and Berk (1980, 1984).

The follow up studies largely fall into two categories. The first category includes research focusing on validating or extending the existing methods. For example, Swaminathan et al. (1974) followed up Hambleton and Novick (1973) and put forward the kappa coefficeint, and Peng and Subkoviak (1980) provided a simple normal approximation to the beta-binomial distributions in Huyhn (1976). The second category consists of the research that examines the impact on the accuracy of DC and DA of factors such as sample size, test length, the number of proficiency-levels, and the placement of the cut-off scores. Ercikan and Julian (2002) represent such an effort. Their paper examined the degree of variation in the accuracy of classifying student performance to proficiency-level scores with changes in the number of proficiency levels and the measurement accuracy. They also examined the degree that the classification accuracy varies across different ability levels given different numbers of proficiency levels based on the same test and the same set of cut-scores. The results of the study indicate that (1) the classification accuracy decreased, on average, by 10% for an increase of 1 proficiency level, 20% for an increase of 2 proficiency levels, and 20% to 30%for an increase of 3 proficiency levels, (2) classification accuracy varied 10% to 20% for tests with reliabilities that ranged between 0.70 and 0.93, and (3) it is more desirable that classification accuracy be reported separately for different score ranges, particularly for a set of critical score ranges in which classification accuracy seems to be the lowest. But the authors noted at the end of the paper that more research was needed to make their findings general.

Another important topic covered in the literature is comparing the performance of the existing methods. The research on reviewing the literature can also be considered as addressing this topic. Subkoviak (1978) compared the accuracy of the DC estimates across four different methods that had emerged by then: Huynh, 1976; Marshall & Haertel, 1976; Subkoviak, 1976; and Swaminathan et al., 1974. The data he used consisted of the responses of 1586 students to parallel forms of 10, 30, and 50 items each from the Scholastic Aptitude Test (SAT). The proportion of consistent classifications on two tests for a population of 1586 was regarded as the criterion and the DC estimated from the four methods from three different samples were compared against it.

All four procedures appear to provide reasonably accurate estimates of DC, for the various cases considered. However, relative advantages and disadvantages can be observed. The Swaminathan et al. procedure produces unbiased estimates; but it requires two test administrations and standard errors are relatively large for classroom size samples. The Huynh, MarshalI-Haertel, and Subkoviak approaches require only one administration of the test and standard errors of estimate are relatively small, but for short tests. Each procedure appears prone to a different type of systematic bias. Subkoviak noted that all things considered, the Huynh approach seems worthy of recommendation. It is mathematically sound, requires only one testing, and produces reasonably accurate estimates, which appear to be slightly conservative for short tests.

2.5 Conclusions Based on the Review of Literature

This review of the literature reveals that most of the existing procedures to estimating DA and DC are in the framework of CTT, and that most commonly used CTT methods are based on single test administration but are often complicated mathematically and computationally. As noted by Brennan and Wan, Livingston and Lewis method is one of the commonly used methods nowadays, but the procedure is rather complicated, and the method itself is not well studied.

Actually one of the reasons for the lack of research on this topic might be related to the difficult task of estimating true score distributions for calculating DA and DC. In literature, one way to solve this problem has been to assume an assumption that candidates' true scores follow a certain distribution. Several CTT methods previously mentioned make such assumptions and often complicated mathematically. In general, there have been several methods for estimate true scores, but it seems none of them can make things any simpler. For example, calculating plausible values can give us some kind of student ability estimates. Plausible values were first developed for the analyses of NAEP data in the 1980's. Because the design of this large-scale test has to accommodate a wide range of content representation by aggregating responses across all respondents, it can only solicit few responses from each sampled respondent. This makes it hard to give precise estimation of each respondent's ability score and therefore the population distribution characteristics. Plausible values methodology was developed as a way to address this issue by using all available data to estimate directly the characteristics of student populations and subpopulations, and then generate imputed scores or plausible values from these distributions that can be used in analyses with standard statistical software (Mislevy, 1991). In doing so, extra data are needed, and the values themselves are not appropriate to be used as individual student scores. Brennan and Wan (2004) also developed a method to calculate DC with the technique bootstrapping. They noted that by using their method, no assumptions of true score distributions needs to be made. This is a great advantage of their method, but their results also indicated that more research is needed for this method to be used widely. Besides, they noted that although it is mathematically simple, bootstrapping can be computationally intensive and demands more computation time than some other methods.

Also, although IRT has been the mainstay in the field of educational and psychological measurement, there is a lack of research and application of evaluating DA and DC in the framework of IRT, and there is a lack of research on comparing the performance of the existing methods, which leads to the absence of clear guidelines as to which method to choose for test users. These findings show that more work in this area is warranted.

The primary motivation of this project is to explore approaches to estimating DA and DC under IRT. While producing robust estimates of DA and DC indices in IRT is certainly the principal objective, the concern that some methods can be computationally intensive will certainly be taken into consideration. Comparing the performance of the most commonly used CTT methods and the IRT methods that will be introduced in the project constitutes another main focus for this study. Finally, software has been written to implement all the procedures under study, and the software for the IRT methods have proven to be computationally efficient, essentially taking no more computing time than the most commonly used CTT methods.

The primary goal of this chapter is to explore approaches to estimating DA and DC in IRT. The method and design of this project consists of three studies: Chapter 3 focuses on the Rudner method and its extensions. It proposed an extension of Rudner's (2001, 2004) method to the evaluation of DC, proposed a new set of methods for calculating DA and DC on the test score metric, and illustrated how they could be applied in a real life situation. Chapter 4 establishes the procedures for extending and implementing the Hambleton and Han (2004) method to polytomous data. Chapter 5 focuses on the comparisons of the commonly used CTT and IRT methods using both real and simulated data. The study ends with the summary of the findings from these 3 studies, implications in relation to the value of these methods, and the identification of more aspects of research that could be done with the IRT methods.

Table 2.3.1 Summary of DC and DA estimation methods*

Method	**One-admin.**	**Two-admin.**	**0-1 Data**	**0-m Data**	**CTT-Based**	**IRT-Based**
Hambleton & Novick (1973)		√	√		√	
Swaminathan, Hambleton & Algina (1974)		√	√		√	
Swaminathan, Hambleton & Algina (1975)		√	√		√	
Subkoviak (1976)	√		√		√	
Huynh (1976)	√		√		√	
Livingston & Lewis (1995)	√		√	√	√	
Rudner (2001)**	√		√			√
Rudner (2004)**	√		√	√		√
Hambleton & Han (2004)	√		√	√		√

Note: * Adapted from Hambleton and Li (2005).

**Ruder methods are for DA estimates only.

Table 2.3.6.1.1 Contingency table for calculating DA from Rudner's method

$\hat{\theta}$ \ θ	$[\theta_{c_0},\theta_{c_1})$	$\cdots$	$[\theta_{c_{i-1}},\theta_{c_i})$	$\cdots$	$[\theta_{c_{k-1}},\theta_{c_k}]$
$[\theta_{c_0},\theta_{c_1})$	p_{11}	$\cdots$	p_{1i}	$\cdots$	p_{1k}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$[\theta_{c_{i-1}},\theta_{c_i})$	p_{i1}	$\cdots$	p_{ii}	$\cdots$	p_{ik}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$[\theta_{c_{k-1}},\theta_{c_k}]$	p_{k1}	$\cdots$	p_{ki}	$\cdots$	p_{kk}

		Classifications based on Test Two	
		Master (+)	Non-Master (-)
Classifications based on Test One	Master (+)	Consistent (+, +)	Inconsistent (+, -)
	Non-Master (-)	Inconsistent (-, +)	Consistent (-, -)

Figure 2.3.1.1 Decision consistency.

		Classifications based on true scores	
		Master (+)	Non-Master (-)
Classifications based on observed scores	Master (+)	Consistent (+, +)	Inconsistent (+, -)
	Non-Master (-)	Inconsistent (-, +)	Consistent (-, -)

Figure 2.3.1.2 Decision accuracy.

CHAPTER 3

EXTENSIONS TO THE RUDNER METHOD

This chapter deals with the Rudner method and its extensions. Specifically, since the Rudner method was developed for calculating DA only, this chapter first proposes an extension of Rudner's (2001, 2004) method to the evaluation of DC. As has been mentioned, given some of the limitations in applying Rudner's (and its extension) method to DA and DC, this chapter then puts forward a new set of methods for calculating DA and DC on the test score metric. Finally, this chapter ends with a real data example illustrating how the proposed two sets of methods can be applied to estimate DA and DC in a real life situation. For the sake of convenience, the Rudner method will be referred to as RM in the rest of the text.

3.1 Method

3.1.1 Extension of RM to DC

As was mentioned earlier, Rudner focused his attention on estimating DA, but DC is a topic of equal importance and is therefore of equal interest to many researchers. The following paragraphs detail the procedures for extending his method to DC. In order to be clear and coherent, the notation system used by Rudner in his original paper addressing DA was maintained as much as possible.

Recall that Rudner referred to θ and $\hat{\theta}$ as 'true score' and 'observed score', respectively in his papers, and his method is based on the fact that for any given true score θ, the corresponding observed score $\hat{\theta}$ is expected to be normally distributed, with a mean θ and a standard deviation of $se(\hat{\theta})$. Also recall that according to Hambleton and Novick (1973), DC reflects the consistency of classifications across repeated testing. If the examinees are administered one test form, the probability of an examinee with a given true score θ of having an observed score in the interval $[a,b]$(an interval between two cut scores) on the theta scale is given by

$$p(a<\hat{\theta}<b\,|\,\theta)=\phi\left[\frac{b-\theta}{se(\hat{\theta})}\right]-\phi\left[\frac{a-\theta}{se(\hat{\theta})}\right]. \tag{11}$$

If the examinees are given the test again without having acquired any practice effect in between the two administrations of the test, the probability of an examinee with a given true score θ of having an observed score in the interval $[c,d]$(another interval between two cut scores) on the theta can be given by

$$P(c<\hat{\theta}<d\,|\,\theta)=\phi\left[\frac{d-\theta}{se(\hat{\theta})}\right]-\phi\left[\frac{c-\theta}{se(\hat{\theta})}\right]. \tag{12}$$

Since the responses to the two tests are independent, the probability of an examinee with a given true score θ of having an observed score in the interval $[a,b]$on the first administration of the test and in the interval $[c,d]$on the second administration of the test is given by

$$P(a<\hat{\theta}<b\,|\,\theta)*P(c<\hat{\theta}<d\,|\,\theta)=$$
$$\{\phi\left[\frac{b-\theta}{se(\hat{\theta})}\right]-\phi\left[\frac{a-\theta}{se(\hat{\theta})}\right]\}*\{\phi\left[\frac{d-\theta}{se(\hat{\theta})}\right]-\phi\left[\frac{c-\theta}{se(\hat{\theta})}\right]\} \tag{13}$$

By extending this logic to all candidates in the test, or to the entire theta scale range, as formulated in equation 14, we can get the expected proportions of all examinees that have observed scores in the interval $[a,b]$ on one form and observed scores in the interval $[c,d]$ on the other form is:

$$\int_{\theta=-\infty}^{+\infty} P(a<\hat{\theta}<b \mid \theta) * P(c<\hat{\theta}<d \mid \theta) f(\theta) d\theta =$$

$$\int_{\theta=-\infty}^{+\infty} \left\{ \left[\phi\left[\frac{b-\theta}{se(\hat{\theta})}\right] - \phi\left[\frac{a-\theta}{se(\hat{\theta})}\right] \right] * \left[\phi\left[\frac{d-\theta}{se(\hat{\theta})}\right] - \phi\left[\frac{c-\theta}{se(\hat{\theta})}\right] \right] \right\} \Phi(\frac{\theta-\mu_\theta}{\delta_\theta}) d\theta \quad , \qquad (14)$$

where $se(\theta)$ is the reciprocal of the square root of the Test Information Function (TIF) at theta, which is the sum of the Item Information Function, and as in the RM for the consistency accuracy, $f(\theta)$ is the standard normal density function $\Phi(Z)$.

When $a = c$, and $b = d$, the above formula gives the DC index.

Analogous to the DA part, the elements of the contingency table are depicted by the following:

$$P_{ij} = \int_{\theta=-\infty}^{\infty} P(\theta_{c_{i-1}} \le \hat{\theta} < \theta_{c_i}) * P(\theta_{c_{i-1}} \le \hat{\theta} < \theta_{c_i}) f(\theta) d\theta \quad , \qquad (15)$$

where $0<i<k, 0<j<k.$ When $i = j$, we have the following formula:

$$P_{ii} = \int_{\theta=-\infty}^{\infty} P(\theta_{c_{i-1}} \le \hat{\theta} < \theta_{c_i}) * P(\theta_{c_{i-1}} \le \hat{\theta} < \theta_{c_i}) f(\theta) d\theta \quad . \qquad (16)$$

The overall DC index is calculated as the sum of the diagonal elements in the contingency table.

3.1.2 A Proposed Method on the Theta Metric

3.1.2.1 Decision Accuracy

The RM is based on the assumption that for any given true score θ, the corresponding observed score $\hat{\theta}$ is expected to be normally distributed, with a mean θ and a standard deviation of $se(\hat{\theta})$. However, this fact holds only when $\hat{\theta}$'s are maximum likelihood estimates. Furthermore, in the Rudner method, examinee scores are based on IRT theta estimates. But because of some practical issues, though tests are often developed and equated with IRT, in some states it is not uncommon for the student scores to be reported using either raw scores or some type of transformation of theta that is on the raw score scale, and the performance standards to be set on the raw score scale too. For example, the modified Angoff method, which is one of the two most commonly used standard setting methods, yields results on the test score metric. In this situation, it is more desirable to derive DA using raw scores. Thus, Rudner's method will be modified in this study to function on the test score metric, and henceforth, referred to as the adapted Rudner method (ARM).

The ARM is appropriate for tests that: (1) the decision of which performance category an examinee should be classified into is made based on the number correct score scale, and (2) cut points determining the performance categories for a test are set on the raw score scale rather than on the scale of examinees' theta estimates. It should be noted that in these situations, the RM is still applicable, but we have to first map cut scores and all students' raw scores to the theta metric via the test characteristics curve (TCC).

For a student with a given true score θ, the probability of her/his obtaining a raw score with its values ranging from zero to the total possible score is a compound binomial (for dichotomous items) or compound multinomial distribution (for polytomous items). To be specific, for a student with a given true score, by applying appropriate IRT model(s), we can compute her/his probability associated with getting each score point on each individual item, and then the probability of his/her getting each score point ranging from zero to the total score on the entire test. For each student, summing her/his probabilities associated with each score point within each score category yields the possibility of being classified into each score category based on the observed score. Thus another method for obtaining DA is as follows:

$$\int_c^d P(a' \leq x < b' \mid \theta) f(\theta) d\theta =$$

$$\int_c^d [P(x = a' \mid \theta) + P(x = a' + 1 \mid \theta) + ... + P(x = b' - 1 \mid \theta)] f(\theta) \Phi(\frac{\theta - \mu_\theta}{\delta_\theta}) d\theta \qquad (17)$$

where $P(a' \leq x < b' \mid \theta)$ represents the probability of observing a raw score x in the interval of $[a', b']$ given a particular theta value; $P(x = x_0 \mid \theta)$ $(a' \leq x_0 < b')$ refers to the probability for a given theta of obtaining each score point, and $f(\theta)$ refers to the standard normal density of examinees' theta values (referred to as true scores by Rudner) which is assumed to be normal as examinees' true distribution can never be known.

In (17), $P(x = x_0 \mid \theta)$ can be derived directly as conditional probabilities, or by the extension of the Lord-Wingersky (1984) recursive algorithm for the dichotomous items to polytomous items developed by Hanson (1994) and Thissen, Pommerich, Billeaud, and Williams (1995). Also, to avoid complicated integration, the following discretized form of (17) was used:

$$\int_{c}^{d} P(a' \leq x < b' \mid \theta) f(\theta) d\theta =$$

$$\sum_{\theta=c}^{d} [P(x=a' \mid \theta)+P(x=a'+1 \mid \theta)+...+P(x=b'-1 \mid \theta)] f(\theta) \Delta\theta \quad (18)$$

Let's again consider the case in which we have a test that is measuring a single latent ability θ, and students are divided into k categories based on the test. Also, let c_0 denote the minimum possible score for the test, c_k the maximum possible score for the test, and c_i or c_j (i, $j \neq 0$ or k) the i^{th} cut-off score. Table 3.1.1.2.1 gives the contingency table for the DA indices with the second method which is followed by steps to carry out the analysis operationally.

With this method, the first step in the Rudner method for transforming the cut-off scores from the raw score scale to the θ score scale is no longer necessary. The elements of the contingency table are conditional probabilities that are calculated as follows:

$$P_{ij} = \int_{c_{j-1}}^{c_j} P(c_{i-1} \leq \hat{c} < c_i) f(\theta) d\theta \quad (19)$$

When $i = j$,

$$P_{ii} = \int_{c_{i-1}}^{c_i} P(c_{i-1} \leq \hat{c} < c_i) f(\theta) d\theta \quad (20)$$

For example, when $i = j = 1$,

$$P_{11} = \int_{c_0}^{c_1} P(c_0 \leq \hat{c} < c_1) f(\theta) d\theta \quad (21)$$

3.1.2.2 Decision Consistency

As has been discussed in the previous section on DA, this method operates on the test score metric, and for a student with a given true score, by applying appropriate IRT model(s), we can compute her/his probability associated with getting each score point on each individual item, and then the probability of his/her getting each score point ranging from zero to the total score on the entire test.

For each student, summing her/his probabilities associated with getting each score point within each score category yields the possibility of being classified into the score category $[a',b']$ based on the observed score, which can be represented by the following:

$$P(a' \le x < b' \mid \theta) = P(x = a' \mid \theta) + P(x = a' + 1 \mid \theta) + ... + P(x = b' - 1 \mid \theta). \qquad (22)$$

If the examinees are then administer the test again without having acquired any practice effect, the probability of an examinee with a given true score θ of having an observed score in the interval $[c',d']$ (another interval between two cut scores) on the theta can be given by

$$P(c' \le \hat{\theta} < d' \mid \theta) = P(x = c' \mid \theta) + P(x = c' + 1 \mid \theta) + ... + P(x = d' - 1 \mid \theta). \qquad (23)$$

Since their responses to the two tests are independent, the probability of an examinee with a given true score θ of having an observed score in the interval $[a',b']$ on the first administration of the test and in the interval $[c',d']$ on the second administration of the test is given by

$$P(a' < \hat{\theta} < b' \mid \theta) * P(c' < \hat{\theta} < d' \mid \theta) =$$

$$[P(x = a' \mid \theta) + P(x = a' + 1 \mid \theta) + \ldots + P(x = b' - 1 \mid \theta)] *$$

$$[P(x = c' \mid \theta) + P(x = c' + 1 \mid \theta) + \ldots + P(x = d' - 1 \mid \theta)] \qquad (24)$$

By extending this logic to all candidates in the test, or to the whole theta scale range, as formulated in the following formula, we can get the expected proportions of all examinees that have observed scores in the interval $[a', b']$ on one form and observed scores in the interval $[c', d']$ on the other form, which is formulated by:

$$\int_{\theta=-\infty}^{+\infty} P(a' < \hat{\theta} < b' \mid \theta) * P(c' < \hat{\theta} < d' \mid \theta) f(\theta) d\theta =$$

$$\int_{\theta=-\infty}^{+\infty} [P(x = a' \mid \theta) + P(x = a' + 1 \mid \theta) + \ldots + P(x = b' - 1 \mid \theta)] *$$

$$[P(x = c' \mid \theta) + P(x = c' + 1 \mid \theta) + \ldots + P(x = d' - 1 \mid \theta)] \Phi(\frac{\theta - \mu_\theta}{\delta_\theta}) d\theta \qquad (25)$$

where *f*(θ) is the standard normal density function $\Phi(Z)$. When $a' = c'$, and $b' = d'$, the above formula gives the DC index.

Analogous to its DA part, the elements of the contingency table are depicted by the following:

$$P_{ij} = \int_{\theta=-\infty}^{\infty} P(c_{i-1} \le x < c_i) * P(c_{j-1} \le x < c_j) f(\theta) d\theta \,, \qquad (26)$$

where $0 < i < k$, $0 < j < k$. When $i = j$, the equation reduces to:

$$P_{ii} = \int_{\theta=-\infty}^{\infty} P(c_{i-1} \le c < c_i) * P(c_{i-1} \le x < c_i) f(\theta) d\theta \,. \qquad (27)$$

The overall DC index is calculated as the sum of the diagonal elements in the contingency table.

3.1.3 Description of Test Data

The test adopted in this study was a high stakes grade-level mathematics assessment form a statewide examination program. It consisted of responses from about 20,000 examinees to 40 items, including 35 dichotomous items and 5 polytomous items. For the purpose of this study, 5,000 from the 20,000 examinees were randomly selected and used. Also, as has been mentioned before, this test had three cut-off scores which formed four score categories. In applying the two IRT methods to the real dataset, these cut-off scores were still adopted.

For all item parameter calibrations, PARSCALE 4 (Muraki & Bock, 2003) was used, with the three-parameter logistic (3-PL) model being used for the dichotomous items and the graded response model (GRM) for the polytomous items. The 3-PL model is given by the following formula:

$$P_{ij} = c_i + (1 - c_i)\frac{\exp[a_i(\theta j - b_i)]}{1 + \exp[a_i(\theta j - b_i)]} \tag{28}$$

where θj is the ability of examinee j , b_i the difficulty parameter of item i, a_i the discrimination parameter of item i, c_i the guessing parameter of item i. The GRM is mathematically stated in the following way (notation from class is used here):

$$P^*_{ix}(\theta) = \frac{e^{Dai(\theta - bix)}}{1 + e^{Dai(\theta - bix)}} \tag{29}$$

where P^*_{IX} (θ) is the cumulative category characteristic function, representing the conditional probability with which an examinee of ability θ obtaining a score of x or higher on item *i*. b_{ix} is the category boundary for score x, representing the probability of an examinee with given ability θ has a 50% chance of getting a score of x or higher. a_i is the item discrimination parameter. It varies across items but is held constant over response categories within one item, and D is the scaling factor equal to 1.7.

3.2 Results

3.2.1 Rudner Method

Table 3.2.1.1 and 3.2.1.2 summarize the DA and DC indices yielded from RM respectively. It is worth pointing out that these two tables represent a typical example of how DA and DC of performance classifications are being reported. For example, with respect to Table 3.2.1.1, the rows represent the proportion of examinees in the total sample who were classified into a certain category based on their observed scores, and columns represent the proportion of examinees who were classified into a certain category based on their true scores. Each of the diagonal elements represents the proportion of examinees in the total sample who were consistently classified into a certain category based on both their observed scores and true scores, and summing up all the diagonal elements yields the total DA index. Similarly, take Table 3.2.1.2 for instance, the rows represent the proportion of examinees in the total sample that were classified into a certain category based on their observed scores on two administrations of a test (with the second one being hypothetical). Each of the diagonal elements represents the proportion of examinees in the total sample who were consistently classified into a certain category based on both their observed scores, and summing up all the diagonal elements yields the total DC index.

Table 3.2.1.1 tells us that with this real test, by using RM, the proportions of students who were classified into each of the four categories based on their observed scores were: 0.2847, 0.3629, 0.2435, and 0.0189 respectively, the proportions based on their true scores were: 0. 2860, 0.3675, 0.2417 and 0.1047 respectively, and the total DA index was 81.61%. Regarding DC with RM on this test, Table 3.2.1.2 displays the following information: the proportions of students who were classified into each of the four categories were 0.2847, 0.3629, 0.2435, and 0.1088 respectively for both of the administrations. The total DC was 74.08% and Kappa was 63.80%.

3.2.2 Adapted Rudner Method

Table 3.2.2.1 and Table 3.2.2.2 give the DA and DC resulted from ARM respectively in this real dataset. According to Table 3.2.2.1, the proportions of students classified into each performance category based on their true scores were 0.2860, 0.3675, 0.2417, and 0.1047 respectively, the proportions based on their observed scores were 0.2749, 0.3568, 0.2451, and 0.1230 respectively, and the total DA was 79.82%. Regarding to DC from the ARM, Table 3.2.2.2 shows that the proportions based on their observed scores were 0.2749, 0.3508, 0.2451, and 0.1230 respectively on both administrations, and the total DC was 72.11%, with Kappa being 61.37%.

Table 3.1.1.2.1 Contingency table for calculating DA from the Adapted Rudner method.

θ $\hat{c}$	$[c_0, c_1)$	$\cdots$	$[c_{i-1}, c_i)$	$\cdots$	$[c_{k-1}, c_k]$
$[c_0, c_1)$	P_{11}	$\cdots$	P_{1i}	$\cdots$	P_{1k}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$[c_{i-1}, c_i)$	P_{i1}	$\cdots$	P_{ii}	$\cdots$	P_{ik}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$[c_{k-1}, c_k]$	P_{k1}	$\cdots$	P_{ki}	$\cdots$	P_{kk}

Table 3.2.1.1 DA results: Rudner method (DA=81.61%).

		True Score Category				Total
Observed		1	2	3	4	
Score	1	0.2479	0.0368	0.0000	0.0000	0.2847
Category	2	0.0381	0.2924	0.0323	0.0000	0.3629
	3	0.0000	0.0383	0.1881	0.0171	0.2435
	4	0.0000	0.0000	0.0212	0.0876	0.1089
Total		0.2860	0.3675	0.2417	0.1047	0.9999

Table 3.2.1.2 DC results: Rudner method (DC=74.08%, Kappa=63.80%).

		Observed Score Category				Total
Observed		1	2	3	4	
Score	1	0.2319	0.0527	0.0002	0.0000	0.2847
Category	2	0.0527	0.2604	0.0496	0.0002	0.3629
	3	0.0002	0.0496	0.1668	0.0269	0.2435
	4	0.0000	0.0002	0.0269	0.0817	0.1088
Total		0.2847	0.3629	0.2435	0.1088	0.9999

Table 3.2.2.1 DA results: Adapted Rudner method (DA=79.82%).

		True Score Category				Total
Observed Score Category		1	2	3	4	
	1	0.2403	0.0346	0.0000	0.0000	0.2749
	2	0.0457	0.2835	0.0276	0.0000	0.3568
	3	0.0000	0.0494	0.1827	0.0130	0.2451
	4	0.0000	0.0000	0.0313	0.0917	0.1230
Total		0.2860	0.3675	0.2417	0.1047	0.9999

Table 3.2.2.2 DC results: Adapted Rudner method (DC=72.11%, Kappa=61.37%).

		Observed Score Category				Total
Observed Score Category		1	2	3	4	
	1	0.2191	0.0555	0.0002	0.0000	0.2749
	2	0.0555	0.2481	0.0496	0.0004	0.3568
	3	0.0003	0.0528	0.1668	0.0304	0.2451
	4	0.0000	0.0004	0.0269	0.0922	0.1230
Total		0.2749	0.3568	0.2451	0.1230	0.9999

CHAPTER 4

EXTENSIONS TO THE HAMBLETON AND HAN METHOD

The previous chapter described explicitly the procedures of two IRT methods for estimating DA and DC. As has been mentioned in the literature review, Hambleton and Han (will be referred to as H&H) represents another approach to evaluating DA and DC in the framework of IRT. This chapter detailed the procedures for extending this method to the tests that contain polytomous data, and illustrated how this method (and its extensions) can be applied in real life situation with a real data set.

4.1 Method

4.1.1 Extension of H&H to Tests Including Polytomous Items

Single administration estimates of DA and DC for polytomous response data features an emerging request from testing agencies and practitioners. As was mentioned in the literature review section, required by their specific task, Hambleton and Han figured out their method to estimate DA and DC for 0-1 data, but the extension of this method to polytomous response data or mixed format data is necessary and the process to do so is straightforward. To be specific, when polytomous items are involved, their three-step IRT-based method for evaluating single administration DA and DC for tests including dichotomous items can still be followed, with the only difference lying in Step 1, and specifically, in the procedures of simulating the item score (or response matrix data) for the examinees.

Recall that Step 1 of their method is the estimation of the ability score distribution and the generation of the observed scores, which entails the following three procedures: (1) A user is often provided with a response data matrix of the examinees so that he/she can proceed with calibrating the response data matrix by applying appropriate IRT model(s) and software to obtain the item parameter and ability estimates; (2) Kelley regressed estimates of ability, which are derived by applying Equation 10 to the ability estimates derived in (1), are substituted for ability score estimates; and (3) For each regressed estimate of ability, based on the assumption that the model and associated parameter estimates are true, both item scores (response matrix) and a total test score are generated for each test.

In the third procedure described above, a routine Monte-Carlo simulation is used to generate the item scores. As has been explained before, with dichotomous items in this process, a probability *p* is to compare with a [0, 1] uniform distribution random number *r*. If $p > r$, the event with the probability *p* happens. Otherwise the event does not happen. An event that happened means the examinee scores 1 on the item, and an event that did not happen means the examinee scores 0 on the item.

But with polytomous items, the simulation process is comparatively more complicated. Suppose the probabilities on each score category for a polytomous item are: $p_0, p_1, \cdots, p_m$. In the polytomous item situation, a [0, 1] uniform distribution random number r is compared with $p'_0 = p_0, p'_1 = p_0 + p_1, \cdots, p'_m = p_0 + p_1 + \cdots + p_m$, ($p'_m = 1$) }; If $r < p'_0$, the examinee scores 0; If $p'_j < r < p'_{j+1}$, the examinee scores *j+1* (Hambleton & Han, in press). By following the same logic, a simulated response matrix on a second test for the same examinee group can also be obtained, and Step 2 and Step 3 of Hambleton and Han method can be followed to calculate DC, kappa, and DA indices.

4.2 Applying of H&H to Real Data

4.2.1 Description of Data Used

The same test used in Chapter 3 was adopted in this study, as were the performance categories and the related IRT models. This allows us to examine the consistency of results derived from these IRT methods, which can also serve as part of the effort to evaluate the performance of these methods.

4.2.2 Results

The DA and DC indices resulting from applying the H&H method were displayed in Table 4.2.2.1 and Table 4.2.2.2 respectively. According to Table 4.2.2.1, the proportions of students who were classified into each of the four categories based on their true scores were 0.2996, 0.3728, 0.2504, and 0.0772 respectively, the proportions based on their observed scores were 0.3096, 0.3848, 0.2486 and 0.0570 respectively, and the total DA index was 79.68%. Table 4.2.2.2 shows us that the proportions of students classified into each of the four categories based on one administration were 0.2996, 0.3728, 0.2504, and 0.0772 respectively, the proportions based on another administration were 0.3024, 0.3714, 0.2544 and 0.0718 respectively, and the total DC index was 71.62%, with Kappa being 58.93%.

Given that Hambleton and Han method produced DA and DC indices that were close to those from both Rudner and Adapted Rudner methods, it seemed the three methods yielded consistent results. However, since these three methods were all pretty new, no evaluation studies have been carried out to evaluate their performance. Therefore, evaluating these methods with simulation studies and examining their robustness in various reasonable conditions should be a crucial step before any sound conclusions can be drawn on their validity and feasibility, and this important step was carried out in the next chapter.

Table 4.2.2.1 DA results: Hambleton and Han method (DA=79.68%).

		True Score Category				Total
Observed Score Category		1	2	3	4	
	1	0.2634	0.0462	0.0000	0.0000	0.3096
	2	0.0362	0.2968	0.0516	0.0002	0.3848
	3	0.0000	0.0298	0.1892	0.0296	0.2486
	4	0.0000	0.0000	0.0096	0.0474	0.0570
Total		0.2996	0.3728	0.2504	0.0772	1.0000

Table 4.2.2.2 DC results: Hambleton and Han method (DC=71.62%, Kappa=58.93%).

		Observed Score Category				Total
		1	2	3	4	
Observed Score Category	1	0.2422	0.0598	0.0004	0.0000	0.3024
	2	0.0572	0.2596	0.0546	0.0000	0.3714
	3	0.0002	0.0530	0.1692	0.0320	0.2544
	4	0.0000	0.0004	0.0262	0.0452	0.0718
Total		0.2996	0.3728	0.2504	0.0772	1.0000

CHAPTER 5

INVESTIGATING THE ROBUSTNESS OF THE IRT METHODS

The previous two chapters have explicitly described the procedures of three IRT methods for estimating DA and DC. The purpose of this study is to present a series of simulation studies under a variety of reasonable conditions designed to examine and compare the robustness of these IRT methods. In these simulated conditions, the observed DA and DC values resulted from these IRT methods were compared to “true” DA and DC values to check on their accuracy, and to the results among themselves to check on their consistency. In addition, the three IRT methods were compared with the commonly used CTT method – the Livingston and Lewis (L&L) method – to see if the IRT methods function at least the same if not more accurately and efficiently. Detailed descriptions of the simulated studies carried out and their corresponding results are treated in the following two sections of this chapter.

5.1 Method

The design of the study was manipulated to examine the impact of three variables expected to influence the overall precision of the DA and DC calculations: (1) estimation methods, (2) test length, and (3) proficiency distribution shape. The effects of each of these factors on DA and DC values are examined by varying these factors separately or jointly. The following is a discussion of the simulation studies used to optimally examine the robustness of the methods involved, including the various simulation study conditions, data generation, and the evaluation criteria.

5.1.1 Test Length

Five different test lengths were considered with respect to the robustness of all four methods involved in estimating DA and DC. It was expected that the four methods are more likely to show differences with short tests, so the number of items for each test length condition was: (1) 10 items, (2) 20 items, (3) 40 items, (4) 60 items, and (5) 80 items.

All the simulation studies in this chapter were based on a statewide mandatory mathematics assessment for a certain grade. The original assessment itself was composed of 35 dichotomous items (scored 0 to 1) and 5 polytomous items (scored 0 to 4), and the sample size was about 20,000. For each of the conditions carried out, the sample size consists of 5,000 randomly drawn examinees (N=5,000).

All these 40 dichotomous and polytomous items in the original assessment were used to create all test forms at the five levels of test length, and the five test length levels were formed in a nested manner, with each successively longer test encompassing all the items from the previous shorter test length plus 10 or 20 additional items. Table 5.1.1.1 gives the specific number of dichotomous items and polytomous items and total raw score for each test length respectively. Also, for the test forms that contained more than 5 polytomous items, the deficit was made up by randomly drawing the corresponding number of polytomous items from the original 5 ones and repeating them.

5.1.2 Distribution Shape of Ability Scores

The shape of ability distribution was also expected to have some impact on the precision of overall estimation of DA and DC for all methods concerned; therefore, 5 distribution shapes were simulated in this study. Other than the (close to) normal distribution, the beta distribution function was defined by specifying different parameter values of a and b to form the other four distribution shapes that might possibly be encountered in practice: two different degrees of positively skewed unimodal distribution, and two different degrees of the negatively skewed unimodal distribution. The different values considered for a and b parameters are summarized in Table 5.2.1.1.

Graphically, Figure 5.2.1.1 gives a general picture of the two different degrees of the negatively skewed unimodal distribution, and Figure 5.2.1.2 illustrates the two different degrees of positively skewed unimodal distribution. For the sake of convenience, based on the a and b parameter values adopted in the beta function to produce each of these shapes, the terms Shape 26, Shape 46, Shape 0, Shape 64, and Shape 62 were used to refer to the positively skewed , slightly positively skewed, normal, slightly negatively skewed, and negatively skewed distributions in all the plots, respectively. By examining Figure 5.2.1.1 and b, we can see that the negatively skewed distribution is the most skewed one among the 4 degrees of skewedness. Also, the reliability estimates for all the simulated test forms across all these test lengths and distribution shapes were given in Table 5.1.2.2. Generally, all the 10-item tests had a reliability value of around 0.70 or less, the 20-item tests around 0.83, the 40-items around 0.91, the 60-item tests around 0.94, and the 80-item tests around 0.96.

5.1.3 Correction of Theta Distribution

This condition applies to H&H method only. Since with their method, the authors held that the estimated theta values (theta parameter estimates calibrated with an

IRT software such as PARSCALE adopted in the present study) are more heterogeneous than the distribution of their interest, this distribution of theta estimates was corrected with Kelly regression method as indicated by formula (10). It seems to be a comparatively easy and efficient way to approximate theta distribution. However, in order to validate their way of estimating theta distribution, a condition considered in this chapter is to examine the impact of this correction on the overall precision of DA and DC estimation. Therefore, with H&H only, the results were compared to those derived from H&H method but without applying the Kelly regression on theta estimates.

5.1.4 Summary of Conditions

In total, 125 conditions were considered for this chapter: 5 DA and DC estimation methods (3 IRT, L&L, and the method to generate "true" DA, DC values) by 5 test length levels by 5 distribution shapes, plus 25 comparisons for the H&H method (5 test length levels by 5 distribution shapes), and this yielded 5*5*5+25 = 150 cases handled in this chapter. The analyses for the 150 conditions were carried out independently. For the analysis of each condition, the sample size of 5,000 (N=5,000) was adopted, and whenever appropriate, the distribution of theta estimates was rescaled to have a mean of zero and a standard deviation of one to ensure scale comparability.

5.1.5 Evaluation Criteria

The primary interest of this chapter is to examine the robustness of the IRT methods for estimating DA and DC. As has been mentioned before, since this field was rather understudied, and most of the work concerning these methods or even some methods themselves were rather new in some sense, the performance of these methods should be brought under scrutiny. That is, in our situation, it is highly desirable to make as many comparisons as appropriate based on the results from the simulation studies, and to gather as much evidence as possible in supporting the adoption of these methods in practice.

5.1.5.1 Comparison with "True" DA and DC

To evaluate the three methods, we could apply the methods to real data and examine the consistency of their results. However, the consistent results may be different from "true" DA and DC results, were the true results known. The availability of true DA and DC results suggests the need for a simulation study. As part of the study, simulation studies were conducted to generate "true" DA and DC indices so as to determine the robustness of the these IRT methods. Described below were the specific evaluation procedures adopted to generate "true" DA and DC values for all the test forms.

(1) True ability scores. True item and person parameters were required for the simulation study. In order for them to be realistic so that they can capture some of the important features of real data, these true item and person parameters were not made up. Rather, the existing parameter estimates were used and treated them as if they were true. For this reason, the PARSCALE calibration results (item and person parameter estimates) of a single 40-item statewide compulsory mathematics test with a sample of about 5,000 examinees were used as true item and ability parameters.

(2) Observed ability scores. Based on these true parameter and theta values, a response matrix for 5,000 simulated examinees which has exactly the same number of dichotomous and polytomous items were generated by following the traditional Monte Carlo procedures (some information about how to follow the Monte Carlo procedures

was also described in the introduction to H&H method in literature review). By calibrating the new response matrix obtained with PARSCALE, a new set of parameter and ability estimates were derived, and these ability estimates were treated as the observed scores for the simulated examinees.

(3) Cut scores. The contractor for this statewide mandatory test provided a set of three cut scores on the raw score scale that classify the students into four performance categories. The TCC for the test were calculated and employed to map these cut scores from the raw score scale to the theta scale.

(4) *True* DA index. The number of students classified into each performance category based on both of their true scores and observed scores were counted, and the proportion of those who are consistently classified into the same performance categories gives the True DA index. Recall that in this situation, students' true scores refer to students' ability derived in step (1), and their observed scores refer to their ability estimates derived in step (2). It should be noted that in real life situations, we never have the luxury of knowing students' true scores, and this is why the RM, the ARM or the H&H method has to be used.

True DC index. Carrying out Step (2) twice independently yields two sets of observed scores for two independently simulated tests. Applying cut scores derived in Step (3) on these two test forms, and then the number of students classified into each performance category based on both sets of observed scores was counted. The proportion of those who are consistently classified into the same category on the basis of the two sets of observed scores represents the True DC index.

(5) Regarding how to count the proportion of students who are consistently classified into the same performance categories based on either their true score and observed score (for DA) or based on two sets of their observed scores (for DC) mentioned in step (4), let us see an example for DA (and the method for DC is analogous to that for DA). Consider the case in which we have a test that is measuring a single latent ability θ, and students are divided into k categories based on the test. Also, let c_0 denote the minimum possible score for the test, c_k the maximum possible score for the test, and c_i $(i \neq 0 \text{ or } k)$ the ith cut-off score. Table 5.1.5.1.1 is the contingency table for calculating the true DA indices.

For each student, if he or she is classified into category i based on his or her true score, and he or she is classified into category j based on his or her observed score, we add the number 1 to that category. That is, we have $f_{ij} = f_{ij} + 1$. The value in each cell is represented by $p_{ij} = \frac{f_{ij}}{N}$, where N is the total number of students in the sample. The overall true DA index is the sum of the diagonal elements in the contingency table.

(6) *Observed* DA and DC indices and evaluation of the method: The RM (and its extension to DC), ARM, or H&H was then applied to the simulated response matrix data respectively to derive the observed DA and DC values derived on theta metric, which were compared with their corresponding "true" DA and DC values.

(7) Evaluation criterion: The degree to which the DA and DC indices resulting from the estimation methods correspond to the "true" DA and DC values as checked by the Absolute Difference (AD) for each comparison using the following equation:

$$AD = | C_i - S_i |, \qquad (29)$$

where C_i is the "true" DA or DC value and S_i is the DA or DC value derived from applying one of the estimation methods. The magnitude of the absolute differences between the true and the observed DA and DC indices will be used to judge the relative merits of all procedures.

5.1.5.2 Comparison with L&L Method

Another question needs to be addressed in this study was: why is it necessary to develop these methods? Though this question was answered in previous chapters in light of measurement theories, statistically, do these methods function any better or at least as well as the commonly used classical-theory-based method – Livingston and Lewis method? Therefore, the second criterion adopted to evaluate the IRT methods was to compare the results yielded from these IRT methods with those from the Livingston and Lewis method at all simulated conditions.

5.1.5.3 Reasonableness of the Methods

The reasonableness of the results produced from each of these methods was also examined. For example, questions like these can be asked: (a) Are DC values resulted from these methods smaller than DA values? (b) Are the derived Kappa values smaller than DC values? (3) Will these methods be computational prohibitive to users in terms of time? and (4) will there be user-friendly software for the users to resort to?

5.2 Results

This section contains the results for the simulation study. As was mentioned previously, three factors were examined in the design of this study: (1) estimation methods, (2) test length, and (3) ability distribution shapes. The five estimation methods (3 IRT methods, one CTT method, and generated "true" values) were completely crossed with the 5 test length levels and the 5 ability distribution shapes, which resulted in 125 conditions. Table 5.2.1, Table 5.2.2 and Table 5.2.3 summarize the overall DA, DC and Kappa values under all these 125 conditions respectively. Two noticeable patterns can be observed by taking a look at these three tables: (1) For all test forms, the results from all methods followed the pattern that DC values were always smaller than DA values, and Kappa values were always less than DC values, as would be expected. (2) Regardless the shape of true score distribution for the test forms, DA, DC and Kappa values all increased as the test length increased. These results were analyzed in detail in the following sections.

Also, across all distribution shapes and test length levels, the H&H method was carried out again by removing the approximation of the ability distribution by applying Kelly regression to the observed theta values. This was crossed with 5 test lengths and resulted in another 5 conditions. Results for these 5 conditions were presented at the end of the results section.

5.2.1 Comparison with "True" DA and DC

The goal of this section is to evaluate the performance of the IRT methods by comparing the results from these methods to "true" DA and DC indices; therefore, the primary outcome of interest in this section is to examine whether or not a particular IRT method provided DA and DC results that were close to the "true" DA and DC values generated in the study, as reflected by the AD values. Table 5.2.1.1.1, Table 5.2.1.1.2 and Table 5.2.1.1.3 display the AD values between the 4 methods and the "true" values for the overall DA, DC and Kappa indices respectively across all 5 test length levels and all 5 true score distribution shapes. In all these tables, the AD values that were larger than 5% were all highlighted by making them in bold. In Figures 5.2.1.1.1 to 5.2.1.1.3, Figures 5.2.1.2.1 to 5.2.1.2.3, Figures 5.2.1.3.1 to 5.2.1.3.3, and Figures 5.2.1.4.1 to 5.2.1.4.3, all AD values were also plotted as the function of test length for the four methods respectively.

The values in these three tables and figures seem to suggest that these four methods are all promising in evaluating DA and DC, since all of them produced results that were close to true values in most of the conditions, as evidenced by the reasonably small AD values. To be exact, most of the AD values across all test length levels and true score distribution shapes as presented in these three tables were within 3%. Also, for every one of the four estimation methods, if it had relatively big AD's on some of the test forms, it always happened in one of the extreme situations: the shortest test, the skewed distribution, or both. Finally, the RM and ARM methods seemed to have fewer larger AD values than the H&H, so did L&L.

By referring to Tables 5.2.1.1.1 to 5.2.1.1.3 and Figures 5.2.1.1.1 to 5.2.1.1.3, we can see that with RM, for DA, all AD values were very small, with only 1 being greater than 3%: 0.0339 (for the positively skewed, 10-item test), and the rest were all within 1%. For DC, 5 out of the 25 AD values were larger than 3%: 0.0352 (for the positively skewed, 10-item test), 0.0315 (for the positively skewed, 20-item test), 0.0399 (for the positively skewed, 60-item test), 0.0322 (for the negatively skewed, 40-item test) and 0.03746 (for the negatively skewed, 80-item test), and the rest were all within 3%. As to Kappa, 2 out of 25 were greater than 5%: 0.0856 (for the positively skewed, 10-item test) and 0.0551 (for the negatively skewed, 80-item test), 4 were between 4% and 5%: 0.0443 (for the negatively skewed, 10), 0.0488 (for the negatively skewed, 40), and 0.0469 (for the negatively skewed, 60-item test), and 3 between 3% and 4%: 0.0375 (the positively skewed, 60-item test), 0.0351 (for the slightly negatively skewed, 40-item test) and 0.0391 (for the slightly positively skewed, 10-item test). The rest were all within 3%.

Tables 5.2.1.1.1 to 5.2.1.1.3 and Figures 5.2.1.2.1 to 5.2.1.2.3 show that with DA values for ARM, all greater than 5% values happened on the10-item test. To be specific, 3 out of the 25 DA values were greater than 5%: 0.0578 (for the slightly positively skewed, 10-item test), 0.0528 (for the normal, 10-item test), and 0.0578 (for the slightly negatively skewed, 10-item test), 1 between 4% and 5%: 0.0426 (for the negatively skewed, 10-item test), and the rest were all smaller than 3%. For DC, 1 was greater than 4%: 0.0461 (for the positively skewed, 60-item test), and the rest were all within 2%. As to AD for Kappa, 2 were greater than 5%: 0.0781 (for the positively skewed, 10-item test), and 0.0519 (for the positively skewed, 60-item test), 2 were between 3% and 4%: 0.0336 (for the negatively skewed, 10-item test) and 0.0367 (for the negatively skewed, 80-item) test.

As shown in Tables 5.2.1.1.1 to 5.2.1.1.3 and Figures 5.2.1.3.1 to 5.2.1.3.3, with respect to H&H, all greater than 5% AD values seemed to be on the 10-item and 20-item tests. To be exact, for DA, 7 out of the 10 AD values on the 10-item and 20-item tests were exceeded 5%. The AD values for other test lengths were very small. For DC, 3 values were greater than 5%: 0.0708 (for the normal, 10-item test), 0.0542 (for the positively skewed, 60-item test), and 0.0586 (for the slightly positively skewed, 10-item test), 3 were between 4% and 5%: 0.0404 (for the negatively skewed, 10-item test), 0.0494 (for the slightly positively skewed, 10-item test), and 0.0422 (for the slightly negatively skewed, 20-item test), 1 was between 3% and 4%: 0.0348 (for the negatively skewed shape, 20-item test), and the rest were within 3%. For Kappa, all 10 AD values for the 10-item and 20-item tests were greater than 5%, and AD for the positively skewed, 60-item test was 0.0676. The rest were all within 3%.

As to L&L, Tables 5.2.1.1.1 to 5.2.1.1.3 and Figures 5.2.1.4.1 to 5.2.1.4.3 indicate that as to DA, all greater than 5% AD values seemed to be on the 10-item and 20-item tests. On 10 of these short test forms, 8 had the AD values larger than 5%. On the other comparatively long test forms, the AD values dropped significantly and most of them were within 3%. Regarding DC, 2 AD values exceeded 5%: 0.0514 (for the slightly positively skewed, 20-item test), and 0.0539 (for the negatively skewed, 20-item test). With respect to Kappa, all AD values for the 20-item tests (except for the negatively skewed one) were larger than 5%.

5.2.2 Comparison with L&L

One of the research questions of this project is to compare all three IRT methods with Livingston and Lewis method (L&L), which also represents a method of checking the consistency of the results they produce. Presented in this section are the comparison of the performance of three IRT methods with that of the L&L method on the test forms of various test lengths and ability distribution shapes. For a general view of their performances, the resultant DA, DC and Kappa values from 3 IRT methods and L&L were plotted in Figures 5.2.2.1.1 to 5.2.2.1.5, Figures 5.2.2.2.1 to 5.2.2.2.5 and Figures 5.2.2.3.1 to 5.2.2.3.5 respectively as a function of test length for each distribution shape.

Their specific differences were also calculated and presented in Tables 5.2.2.1.1 to 5.2.2.1.3. By examining these tables and figures, we can make the following general observations: (1) All 3 IRT methods produced results that were close to those from L&L, especially on the long test forms. This is shown by the fact that in all figures mentioned above, the 4 lines representing the 4 methods were pretty tight and followed the same pattern. (2) Among the 4 methods, RM seemed to produce the highest DA, DC and Kappa values, L&L the lowest, and ARM and H&H were in between, with the values from ARM slightly higher than H&H. In all figures for DA, DC and Kappa, the lines for RM were always on the top of the bundle, the L&L in the bottom, with the lines for ARM and H&H falling in between them. (3) As the test grew longer, the difference between them grew smaller, especially beginning from the 40-item test. In each of the figures mentioned above, it is pretty noticeable that all lines began to grow tighter when the test length was 40. (4) The results from the four methods displayed the most variability on the negatively skewed and 10-item test. Figure 5.2.2.1.5, Figure 5.2.2.2.5 and Figure 5.2.2.3.5 present the resultant DA, DC and Kappa values from the four methods for the tests that have negatively skewed shapes respectively. In each one of them, compared with their other corresponding figures, the 4 lines were the most dispersed at the 10 item length level.

The differences between the results from each of the 3 IRT methods and the L&L were also graphically displayed in Figures 5.2.2.4.1 to 5.2.2.4.3 for DA and 5.2.2.5.1 to 5.2.2.5.3 for DC and 5.2.2.6.1 to 5.2.2.6.3 for Kappa. The observations made based on these figures support those made in the previous paragraph.

On DA, among the three IRT methods, RM had the biggest differences with L&L, and these big differences mainly occurred on the 10-item tests that had the most skewed true score distribution shapes (62 and 26). Also, these differences reduced markedly when the tests grew to 20 items, and reduced even further as the length of the tests grew. ARM had no greater than 5% difference value with L&L, and the negatively skewed and the positively tests displayed the greatest differences. H&H had no greater than 5% differences either. The difference between all IRT methods and L&L were reduced as the test length increased, and for all of them, the negatively skewed tests seemed to have the biggest differences. The tests that had normal true score distributions produced the closest results.

The patterns for DC were similar to those on DA. On DC, RM still had the biggest differences with L&L, as illustrated by Figures 5.2.2.5.1 to 5.2.2.5.3. It had the most greater than 5% difference values with L&L on the positively skewed, 10- and 20-item tests and on the negatively skewed 10- and 20-item tests. The values for the rest were small and reasonable. ARM had 2 greater or close to 5% values on the positively skewed, 10-item test and the negatively skewed, 20-item test. H&H produced the closest results to L&L overall. It had greater variability on short tests and tended to yield smaller DA values for these tests. Again, for all of them, the differences mostly happened on the positively/negatively skewed and the short tests. The test forms that had normal true score distributions yielded the closest results.

Figures 5.2.2.6.1 to 5.2.2.6.3 show a similar pattern. Differences for all methods and L&L reduced as test length increased. RM had the biggest differences, and the biggest differences were found in both the positively and negatively skewed with both 10 and 20 items. ARM and H&H produced closer results than RM. H&H again had the smallest Kappa values for the short tests, and the tests that had normal true score distributions produced the best results.

5.2.3 Effects of Test Length

We have known from 5.2.1 and 5.2.2 that across all test lengths and distribution shapes, all methods concerned displayed some variability and instability on tests that have fewer items and have skewed true score distribution shapes. In this section, in order to better reveal the effect of test length on estimating of DA, DC and Kappa using all the methods concerned, only the normal distribution shape (Shape 0) was considered across the 5 levels of the test length variable: 10, 20, 40, 60 and 80 items.

Tables 5.2.3.1.1, 5.2.3.1.2 and 5.2.3.1.3 present the overall estimates for DA, DC and Kappa respectively for all these test length levels from RM, ARM, H&H, L&L, and the "true" values. From all these tables and graphs, the following observations can be made.

First, DA, DC and Kappa values are expected to vary for tests of different lengths. Specifically, we expect the resulted DA, DC and Kappa values to get greater in increasing order of test length. This trend can be observed by referring to Table 5.2.3.1.1, Table 5.2.3.1.2 and Table 5.2.3.1.3. For example, DA indices from all methods grow from a certain value between 0.61 to 0.65 for the test containing 10 items, to a certain value between 0.71 to 0.75 for the test containing 20 items, a certain value between 0.80 to 0.83 for the test having 40 items, and a value between 0.83 to 0.84 for the test containing 60 items, and a value between 0.86 to 0.88 for test containing 80 items.

Further, a noticeable feature reflected by this kind of growth is that the growth in DA, DC and Kappa indices from all methods is more drastic from the 10-item test to the 20-item test and from the 20-item test to the 40-item test than the growth from the

40-item to the 60-item or from the 60-item to the 80-item test. For example, for all methods, the DA values increased about 0.1 from the 10-item to the 20item test, about 0.08 from the 20-item to the 40-item test, about 0.03 from the 40-item to the 60-item test, and from 60-item to the 80-item test. The DC values gained about 0.11 from the 10- to 20- item test, about 0.9 from the 20- to 40- item test, about 0.4 from the 40- to 60-item test and about 0.5 from the 60- to 80 item test. The Kappa values grew from about 0.16 from the 10- to 20- item test, about 0.12 from the 20- to 40-item test, about 0.4 from the 40- to 60-item test and about 0.7 from the 60- to the 80-item test. Again, for the growth between the 40- to 60-item test and the 60- to 80-item test in all DA, DC and Kappa values, it can also be seen that the latter one – the growth from the 60- to the 80-item test is slightly more marked than the growth from the 40- to the 60-item test.

Secondly, compared with the set of "true" DA, DC and Kappa values, RM, ARM, H&H and L&L methods produced close DA, DC and Kappa results across the 5 different test lengths. Across all 5 test length levels, for each single method, the AD values between the resulting DA, DC and Kappa values with the "true" values were decreasing as test length was increasing, and the variability of the AD values was also decreasing as the test length was increasing. This is also the case with differences between the 3 IRT methods with L&L for the normal shape. The AD values between the IRT methods and the "true" values for DA, DC and Kappa for the normal shape are also graphically presented in Figures 5.2.3.1.1 to 5.2.3.1.3 respectively. The difference values between the IRT methods and L&L method are presented in Figures 5.2.3.2.1 to 5.2.3.2.3 for references.

Referring to Figures 5.2.3.1.1 to 5.2.3.1.3, we can see that so far as DA is concerned, H&H and L&L produced comparatively greater AD values on the 10-item test. The values dropped to within 5% and were more stable when the test grew to 20 items and became pretty stable after the test was 40 items long. The AD between the results from the RM, ARM, H&H and L&L with the "true" results on the 10-item test were 0.0025, 0.0528, 0.0543 and 0.0495 respectively, on the 20-item test, the AD values were 0.0042, 0.0234, 0.0290 and 0.0424 respectively, on the 40-item test, the AD values were 0.0040, 0.0192, 0.0206 and 0.0342 respectively, on the 60-item test, the AD values based were 0.0022, 0.0143, 0.0149, and 0.0240 respectively, and on the 80-item test were 0.0003, 0.0109, 0.0101 and 0.0211 respectively.

Regarding DC values, almost all of the AD values between each of the RM, ARM, H&H and L&L and the "true" values for all test lengths were all very reasonable. For example, on the 10-item test, they were 0.0140, 0.0101, 0.0708 and 0.0443 respectively, on the 20-item test, they were 0.0101, 0.0098, 0.0290 and 0.0424 respectively, on the 40-item test, 0.0091, 0.0087, 0.0144 and 0.0311 respectively, on the 60-item test, 0.0113, 0.0032, 0.0042 and 0.0190 respectively, and on the 80-item test, they were 0.0172, 0.0037, 0.0052, and 0.0128 respectively. The same is true to Kappa values, though RM and ARM tend to produce slightly greater AD Kappa values. For Kappa, the AD values were 0.0070, 0.0030, 0.1640 and 0.0498 respectively for the 10-item test, 0.0105, 0.0113, 0.0566 and 0.0529 respectively for the 20-item test, 0.0174, 0.0050, 0.0247 and 0.0409 respectively for the 40-item test, 0.0190, 0.0015, 0.0084 and 0.0073 respectively for the 60-item test and 0.0238, 0.0052, 0.0073 and 0.0166 respectively for the 80-item test.

Finally, compared with the commonly used L&L method, the IRT methods RM, ARM, and H&H produced close DA, DC and Kappa results across the 5 different test lengths, with no differences for DA or DC values between any method and L&L being greater than 7%, and no differences for Kappa values greater than 8% on all 5 test length levels. All the big AD numbers appeared on the 10-item and 20-item tests.

Also, across all 5 test length levels, for each single method, the AD values between the resulted DA, DC and Kappa values with the "true" values were decreasing as test length was increasing, and the variability of the AD was also decreasing as the test length was increasing.

5.2.4 Effects of True Score Distribution Shape

In this section, 5 levels of the distribution shape for the true scores were considered as a variable and its effect on the accuracy of RM, ARM, H&H, and L&L were examined. We mainly want to find out how robust the RM and ARM are when they are applied to the examinees that have a skewed ability distribution? What about the H&H and L&L since no literature that has been known of has suggested that such robustness test have been conducted on them? If there are differences, how big are the differences for RM and ARM, as compared with those for the H&H and L&L? In order to best address these questions, the test length of 40 items was adopted as it is closest to the length of the test on which the situation studies in this project were based on.

Tables 5.2.4.1.1 to 5.2.4.1.3 present the overall estimates for DA, DC and Kappa respectively for the tests of five different distribution shapes from RM, ARM, H&H, L&L, and True values. Overall, the DA, DC and Kappa values resulted from all methods across all true score distribution shapes were close to each other and roughly follow the same pattern. For all methods, DC values (most of which were around 0.75) were always smaller than DA values (most of which were around 0.80), and Kappa values (most of which were around 0.65) were always less than DC values, as would be expected.

The DA, DC and Kappa values derived on the 40-item test across all 5 distribution shapes were first compared against the "true" values, and the AD between the methods and the true values were calculated and displayed in Figure 5.2.4.1.1, Figure 5.2.4.1.2 and Figure 5.2.4.1.3 for DA, DC and Kappa respectively.

Overall, the resulted AD values seem to suggest that all 4 methods function well across all true score distribution shapes, since most of the AD values were less than 5%. To be exact, for DA, all 20 AD values (4 methods across 5 distribution shapes) were smaller than 4% – actually only 3 values from L&L were greater than 3%. The same is true for DC. That is, no AD values for any method were greater than 4%. As to Kappa, no AD values exceeded 5% for all methods across all distribution shapes. Also, all the AD values for the normal shape test were very small. They are generally smaller for RM and ARM than L&L and sometimes than H&H results, as would be expected, since the two methods have an assumption that examinees' true score distribution is normally distributed. L&L seemed to produce the biggest AD values.

The DA, DC and Kappa results from the three IRT methods were also compared to those from the L&L method, and the results were summarized in Figure 5.2.4.2.1, Figure 5.2.4.2.2, and Figure 5.2.4.2.3 for the comparisons on DA, DC and Kappa respectively. The results of the analyses showed that in general, the three IRT methods produced pretty close results to the L&L method across the 5 distribution shapes.

Specifically, the following observations can be made: (1) Generally, all three IRT methods seemed to produced greater DA, DC and Kappa values than L&L methods across all 5 distribution shapes, with the exception that H&H sometimes had a 1 smaller DC value, and almost all 5 Kappa values. (2) For all three IRT methods, their DA values had the smallest difference with the L&L ones in comparison with DC and Kappa values. For DA, 14 out of all 15 difference values (3 IRT methods across 5 distribution shapes) were within 4% with 1 exceeding 4% but smaller than 5%, for DC, 14 out of 15 were within 5% with 1 from RM on the negatively skewed test greater than 5%, and for Kappa, the differences were greater for RM and H&H, with both of them having greater than 5% differences with L&L. ARM, however, produced very close results. (3) RM seemed to have the biggest differences with L&L on DA, DC and Kappa, with most difference values for DA, DC and Kappa being around or greater than 3%. ARM produced closer results than RM, with all DA differences being around 2%, DC around 3% and Kappa around 4%. H&H method yielded the closest results to L&L, with DA differences varying between around 1% and 2%, DC differences varying between 1% and 3%, and Kappa values no greater than 2%. (4) By examining Figure 5.2.4.2.1, Figure 5.2.4.2.2, and Figure 5.2.4.2.3, it can be noticed that all 3 IRT methods showed pretty close patterns in their difference with L&L results.

5.2.5 Correction on H&H

As was described in the literature review section, H&H approximated true score distribution by applying Kelly regression to the theta estimates derived from the calibration done by an IRT software. Comparisons were made between the "true" values, H&H with this correction, and the H&H without applying this correction to the theta estimates. The results were graphically summarized in Figures 5.2.5.1.1 to 5.2.5.1.5, and Figures 5.2.5.2.1 to 5.2.5.2.5 for DA and DC respectively. It should be noted that since Kappa largely follow the pattern of DC, so their plots were not presented for the sake of parsimony. In each of these two sets of figures, there are 5 panels, each illustrating one of the 5 distribution shapes.

These Figures show that overall the corrected and uncorrected H&H produced close results to "true" values. Actually their results were very similar to the "true" values on the tests that had 40 or more items for both DA and DC. On the short tests, both of them were slightly smaller than the "true" values, with the uncorrected ones being closer to the "true" values most of the time. The differences between the two methods themselves were also small, as indicated by Tables 5.2.5.1.1 to 5.2.5.1.3 that display these differences on DA, DC and Kappa respectively, as well as Figures 5.2.3.1 to 5.2.3.3. On DA, across all distributions, only the 10-item tests produced differences of about or a bit more than 5%. The differences on the 20-item tests dropped to around 2%, and the differences on the 40- or more item tests were very small. The same was true to the DC values. On Kappa, the differences on the 10-item tests were greater, being close to 15%, those on the 20-items were around 7%, the rest were very small. All the differences decreased as the test increased.

Table 5.1.1.1 Test length by item type.

	Test Length				
Item Numbers	**10**	**20**	**40**	**60**	**80**
Dichotomous	8	16	32	48	64
Polytomous	2	4	8	12	16
Total Raw Score	16	32	64	96	128

Table 5.1.2.1 Simulated distribution shapes considered for ability scores.

Case	*a*	*b*	Shape
1	2	6	Unimodal Positively skewed
2	4	6	Unimodal Positively skewed (slightly)
3	6	4	Unimodal The negatively skewed (slightly)
4	6	2	Unimodal The negatively skewed

Table 5.1.2.2 Summary of reliability estimates for all simulated test forms.

Distribution Shape	Test Length				
	10	**20**	**40**	**60**	**80**
26	0.702	0.836	0.912	0.940	0.955
46	0.697	0.826	0.913	0.941	0.960
0	0.683	0.833	0.913	0.941	0.960
64	0.694	0.832	0.913	0.942	0.955
62	0.674	0.827	0.912	0.940	0.953

Table 5.1.5.1.1 Contingency table for calculating true DA indices.

θ $\hat{\theta}$	$[c_0, c_1)$	$\cdots$	$[c_{i-1}, c_i)$	$\cdots$	$[c_{k-1}, c_k]$
$[c_0, c_1)$	p_{11}	$\cdots$	p_{1i}	$\cdots$	p_{1k}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$[c_{i-1}, c_i)$	p_{i1}	$\cdots$	p_{ii}	$\cdots$	p_{ik}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$[c_{k-1}, c_k]$	p_{k1}	$\cdots$	p_{ki}	$\cdots$	p_{kk}

Table 5.2.1 Overall values for all 125 conditions: DA.

Method	Shape	Test Length				
		10	20	40	60	80
RM	26	0.6961	0.7662	0.8280	0.8435	0.8788
	46	0.6671	0.7629	0.8298	0.8592	0.8788
	0	0.6673	0.7596	0.8268	0.8578	0.8793
	64	0.6700	0.7605	0.8280	0.8597	0.8782
	62	0.6501	0.7520	0.8256	0.8559	0.8771
	Mean	0.6701	0.7602	0.8276	0.8552	0.8784
ARM	26	0.6428	0.7321	0.8087	0.8319	0.8687
	46	0.6156	0.7444	0.8120	0.8464	0.8688
	0	0.6170	0.7404	0.8116	0.8457	0.8681
	64	0.6138	0.7420	0.8118	0.8470	0.8673
	62	0.6014	0.7189	0.8103	0.8444	0.8649
	Mean	0.6181	0.7356	0.8109	0.8431	0.8676
H&H	26	0.6242	0.7066	0.8068	0.8208	0.8616
	46	0.5718	0.7170	0.8068	0.8434	0.8698
	0	0.5679	0.7095	0.8102	0.8451	0.8689
	64	0.5830	0.7084	0.7998	0.8394	0.8570
	62	0.5586	0.7038	0.8070	0.8296	0.8600
	Mean	0.5811	0.7091	0.8061	0.8357	0.8635
L&L	26	0.6282	0.7075	0.7950	0.8485	0.8605
	46	0.6072	0.7069	0.7970	0.8354	0.8585
	0	0.6026	0.7143	0.7966	0.8360	0.8579
	64	0.6056	0.7107	0.8059	0.8304	0.8543
	62	0.5855	0.6842	0.7782	0.8186	0.8439
	Mean	0.6058	0.7047	0.7945	0.8338	0.8550
True	26	0.6622	0.7652	0.8332	0.8516	0.8822
	46	0.6734	0.7632	0.8338	0.8620	0.8834
	0	0.6698	0.7638	0.8308	0.8600	0.8790
	64	0.6716	0.7670	0.8280	0.8588	0.8768
	62	0.6440	0.7444	0.8170	0.8430	0.8654

Table 5.2.2 Overall values for all 125 conditions: DC.

		Test Length				
Method	Shape	10	20	40	60	80
RM	26	0.5870	0.6733	0.7572	0.7791	0.8287
	46	0.5545	0.6696	0.7597	0.8010	0.8288
	0	0.5546	0.6655	0.7557	0.7991	0.8294
	64	0.5584	0.6669	0.7574	0.8018	0.8279
	62	0.5393	0.6569	0.7542	0.7965	0.8264
	Mean	0.5588	0.6664	0.7568	0.7955	0.8282
ARM	26	0.5725	0.6471	0.7372	0.7669	0.8152
	46	0.5301	0.6502	0.7395	0.7849	0.8158
	0	0.5305	0.6456	0.7379	0.7846	0.8159
	64	0.5338	0.6478	0.7390	0.7867	0.8142
	62	0.5202	0.6319	0.7369	0.7822	0.8120
	Mean	0.5374	0.6445	0.7381	0.7811	0.8146
H&H	26	0.5248	0.6126	0.7286	0.7588	0.8040
	46	0.4774	0.6322	0.7282	0.7788	0.8286
	0	0.4698	0.6264	0.7322	0.7836	0.8070
	64	0.4858	0.6052	0.7380	0.7746	0.8050
	62	0.4714	0.6020	0.7274	0.7720	0.8082
	Mean	0.4858	0.6157	0.7309	0.7736	0.8106
L&L	26	0.5177	0.6017	0.7114	0.7862	0.8025
	46	0.5005	0.6038	0.7162	0.7679	0.8002
	0	0.4963	0.6130	0.7155	0.7688	0.7994
	64	0.5011	0.6085	0.7281	0.7612	0.7944
	62	0.4804	0.5829	0.6931	0.7474	0.7807
	Mean	0.4992	0.6020	0.7129	0.7663	0.7954
True	26	0.5518	0.6418	0.7474	0.8130	0.8132
	46	0.5268	0.6552	0.7416	0.7920	0.8240
	0	0.5406	0.6554	0.7466	0.7878	0.8122
	64	0.5372	0.6474	0.7312	0.7970	0.8074
	62	0.5118	0.6368	0.7220	0.7672	0.7890

Table 5.2.3 Overall values for all 125 conditions: Kappa.

Method	Shape	Test Length				
		10	20	40	60	80
RM	26	0.4257	0.5446	0.6608	0.6915	0.7604
	46	0.3849	0.5399	0.6643	0.7217	0.7604
	0	0.3689	0.5233	0.6609	0.7191	0.7586
	64	0.3901	0.5362	0.6612	0.7227	0.7593
	62	0.3642	0.5225	0.6567	0.7155	0.7572
	Mean	0.3868	0.5333	0.6608	0.7141	0.7592
ARM	26	0.4182	0.5127	0.6355	0.6770	0.7417
	46	0.3666	0.5172	0.6398	0.6999	0.7428
	0	0.3585	0.5016	0.6385	0.6987	0.7401
	64	0.3730	0.5140	0.6388	0.7029	0.7414
	62	0.3536	0.4957	0.6352	0.6970	0.7388
	Mean	0.3740	0.5082	0.6376	0.6951	0.7410
H&H	26	0.2487	0.4403	0.6161	0.6613	0.7230
	46	0.2217	0.4646	0.6167	0.6878	0.7588
	0	0.1978	0.4563	0.6188	0.6918	0.7275
	64	0.2397	0.4295	0.6296	0.6822	0.7269
	62	0.1896	0.4244	0.6096	0.6741	0.7289
	Mean	0.2195	0.4430	0.6182	0.6794	0.7330
L&L	26	0.3084	0.4470	0.5979	0.6910	0.7223
	46	0.3169	0.4488	0.6065	0.6756	0.7204
	0	0.3120	0.4600	0.6026	0.6741	0.7183
	64	0.3214	0.4570	0.5966	0.6671	0.7137
	62	0.2915	0.4275	0.5715	0.6437	0.6919
	Mean	0.3100	0.4481	0.5950	0.6703	0.7133
True	26	0.3401	0.4979	0.6462	0.7289	0.7361
	46	0.3459	0.5162	0.6393	0.7088	0.7536
	0	0.3618	0.5128	0.6435	0.7002	0.7348
	64	0.3624	0.5064	0.6261	0.7162	0.7313
	62	0.3199	0.4961	0.6079	0.6686	0.7021

Table 5.2.1.1.1 Absolute differences (AD) for all methods: DA.

Method	Shape	Test Length				
		10	20	40	60	80
RM	26	0.0339	0.0010	0.0052	0.0081	0.0034
	46	0.0063	0.0003	0.0040	0.0028	0.0046
	0	0.0025	0.0042	0.0040	0.0022	0.0003
	64	0.0016	0.0065	0.0000	0.0009	0.0014
	62	0.0061	0.0076	0.0086	0.0129	0.0117
ARM	26	0.0194	0.0331	0.0245	0.0197	0.0135
	46	**0.0578**	0.0188	0.0218	0.0156	0.0146
	0	**0.0528**	0.0234	0.0192	0.0143	0.0109
	64	**0.0578**	0.0250	0.0162	0.0118	0.0095
	62	0.0426	0.0255	0.0067	0.0014	0.0005
H&H	26	0.0380	**0.0586**	0.0264	0.0308	0.0206
	46	**0.1016**	0.0462	0.0270	0.0186	0.0136
	0	**0.1019**	**0.0543**	0.0206	0.0149	0.0101
	64	**0.0886**	**0.0586**	0.0282	0.0194	0.0198
	62	**0.0854**	0.0406	0.0100	0.0134	0.0054
L&L	26	0.0340	**0.0577**	0.0382	0.0031	0.0217
	46	**0.0663**	**0.0563**	0.0368	0.0266	0.0249
	0	**0.0672**	0.0495	0.0342	0.0240	0.0211
	64	**0.0660**	**0.0564**	0.0221	0.0284	0.0225
	62	**0.0585**	**0.0602**	0.0388	0.0244	0.0215

Table 5.2.1.1.2 Absolute differences (AD) for all methods: DC.

Method	Shape	Test Length				
		10	20	40	60	80
RM	26	0.0352	0.0315	0.0098	0.0339	0.0155
	46	0.0277	0.0144	0.0181	0.0090	0.0047
	0	0.0140	0.0101	0.0091	0.0113	0.0172
	64	0.0212	0.0195	0.0262	0.0047	0.0205
	62	0.0275	0.0201	0.0322	0.0293	0.0374
ARM	26	0.0207	0.0053	0.0102	0.0461	0.0020
	46	0.0033	0.0050	0.0021	0.0071	0.0082
	0	0.0101	0.0098	0.0087	0.0032	0.0037
	64	0.0034	0.0004	0.0078	0.0103	0.0068
	62	0.0084	0.0049	0.0149	0.0150	0.0230
H&H	26	0.0270	0.0292	0.0188	**0.0542**	0.0092
	46	0.0494	0.0230	0.0134	0.0132	0.0046
	0	**0.0708**	0.0290	0.0144	0.0042	0.0052
	64	**0.0514**	0.0422	0.0068	0.0224	0.0024
	62	0.0404	0.0348	0.0054	0.0048	0.0192
L&L	26	0.0341	0.0401	0.0360	0.0268	0.0107
	46	0.0263	**0.0514**	0.0254	0.0241	0.0238
	0	0.0443	0.0424	0.0311	0.0190	0.0128
	64	0.0361	0.0389	0.0031	0.0358	0.0130
	62	0.0314	**0.0539**	0.0289	0.0198	0.0083

Table 5.2.1.1.3 Absolute differences (AD) for all methods: Kappa.

Method	Shape	Test Length				
		10	20	40	60	80
RM	26	**0.0856**	0.0467	0.0146	0.0375	0.0243
	46	0.0391	0.0237	0.0251	0.0129	0.0068
	0	0.0070	0.0105	0.0174	0.0190	0.0238
	64	0.0277	0.0298	0.0351	0.0065	0.0279
	62	0.0443	0.0263	0.0488	0.0469	**0.0551**
ARM	26	**0.0781**	0.0147	0.0107	**0.0519**	0.0056
	46	0.0207	0.0009	0.0005	0.0089	0.0108
	0	0.0033	0.0113	0.0050	0.0015	0.0052
	64	0.0107	0.0076	0.0127	0.0133	0.0101
	62	0.0336	0.0004	0.0273	0.0284	0.0367
H&H	26	**0.0914**	**0.0577**	0.0301	**0.0676**	0.0131
	46	**0.1242**	**0.0517**	0.0226	0.0210	0.0052
	0	**0.1640**	**0.0566**	0.0247	0.0084	0.0073
	64	**0.1227**	**0.0769**	0.0036	0.0340	0.0044
	62	**0.1304**	**0.0718**	0.0017	0.0055	0.0268
L&L	26	0.0318	**0.0509**	0.0483	0.0379	0.0138
	46	0.0290	**0.0675**	0.0327	0.0332	0.0332
	0	0.0498	**0.0529**	0.0409	0.0261	0.0166
	64	0.0409	0.0494	0.0295	0.0491	0.0176
	62	0.0285	**0.0687**	0.0364	0.0249	0.0102

Table 5.2.2.1.1 Differences between IRT methods and L&L: DA.

Method	Shape	Test Length				
		10	20	40	60	80
RM	26	0.0680	0.0587	0.0329	-0.0050	0.0183
	46	0.0600	0.0560	0.0327	0.0238	0.0204
	0	0.0647	0.0453	0.0302	0.0218	0.0214
	64	0.0644	0.0498	0.0221	0.0293	0.0239
	62	0.0646	0.0678	0.0473	0.0374	0.0333
ARM	26	0.0146	0.0246	0.0137	-0.0166	0.0082
	46	0.0085	0.0375	0.0149	0.0110	0.0103
	0	0.0144	0.0262	0.0150	0.0097	0.0102
	64	0.0082	0.0314	0.0059	0.0166	0.0130
	62	0.0159	0.0346	0.0321	0.0258	0.0210
H&H	26	-0.0040	-0.0009	0.0118	-0.0277	0.0011
	46	-0.0354	0.0101	0.0098	0.0080	0.0113
	0	-0.0347	-0.0048	0.0136	0.0091	0.0110
	64	-0.0226	-0.0022	-0.0061	0.0090	0.0027
	62	-0.0269	0.0196	0.0288	0.0110	0.0161

Table 5.2.2.1.2 Differences between IRT methods and L&L: DC

Method	Shape	Test Length				
		10	20	40	60	80
RM	26	0.0693	0.0716	0.0458	-0.0071	0.0262
	46	0.0540	0.0658	0.0436	0.0331	0.0286
	0	0.0583	0.0526	0.0402	0.0303	0.0299
	64	0.0574	0.0584	0.0293	0.0406	0.0335
	62	0.0589	0.0740	0.0611	0.0492	0.0457
ARM	26	0.0548	0.0454	0.0258	-0.0194	0.0127
	46	0.0296	0.0464	0.0234	0.0170	0.0156
	0	0.0342	0.0326	0.0224	0.0158	0.0164
	64	0.0327	0.0393	0.0109	0.0255	0.0198
	62	0.0398	0.0491	0.0438	0.0349	0.0313
H&H	26	0.0071	0.0109	0.0172	-0.0274	0.0015
	46	-0.0231	0.0284	0.0120	0.0109	0.0284
	0	-0.0265	0.0134	0.0167	0.0148	0.0076
	64	-0.0153	-0.0033	0.0099	0.0134	0.0106
	62	-0.0090	0.0191	0.0343	0.0246	0.0275

Table 5.2.2.1.3 Differences between IRT methods and L&L: Kappa

Method	Shape	Test Length				
		10	20	40	60	80
RM	26	0.1173	0.0976	0.0629	0.0005	0.0380
	46	0.0681	0.0911	0.0578	0.0461	0.0400
	0	0.0569	0.0634	0.0583	0.0450	0.0404
	64	0.0686	0.0791	0.0646	0.0556	0.0456
	62	0.0728	0.0950	0.0852	0.0718	0.0652
ARM	26	0.0359	-0.0013	0.0085	-0.0045	0.0017
	46	0.0077	-0.0032	-0.0028	0.0028	-0.0174
	0	-0.0152	-0.0182	-0.0004	0.0002	0.0074
	64	-0.0003	0.0061	-0.0043	-0.0047	0.0024
	62	0.0134	0.0083	0.0182	0.0157	0.0135
H&H	26	-0.1284	-0.0947	-0.0588	-0.0485	-0.0357
	46	-0.1846	-0.0864	-0.0500	-0.0390	-0.0126
	0	-0.1812	-0.0769	-0.0365	-0.0253	-0.0287
	64	-0.1706	-0.1206	-0.0247	-0.0416	-0.0277
	62	-0.1645	-0.0884	-0.0142	-0.0232	0.0110

Table 5.2.3.1.1 Overall estimates from Shape 0 and all test lengths: DA.

Method	Test Length				
	10	20	40	60	80
RM	0.6673	0.7596	0.8268	0.8578	0.8793
ARM	0.6700	0.7605	0.8280	0.8597	0.8782
H&H	0.6501	0.7520	0.8256	0.8559	0.8771
L&L	0.6428	0.7321	0.8087	0.8319	0.8687
True	0.6698	0.7638	0.8308	0.8600	0.8790

Table 5.2.3.1.2 Overall estimates from Shape 0 and all test lengths: DC.

Method	Test Length				
	10	20	40	60	80
RM	0.5546	0.6655	0.7557	0.7991	0.8294
ARM	0.5305	0.6456	0.7379	0.7846	0.8159
H&H	0.4698	0.6264	0.7322	0.7836	0.8070
L&L	0.4963	0.6130	0.7155	0.7688	0.7994
True	0.5406	0.6554	0.7466	0.7878	0.8122

Table 5.2.3.1.3 Overall estimates from Shape 0 and all test lengths: Kappa.

Method	Test Length				
	10	20	40	60	80
RM	0.3689	0.5233	0.6609	0.7191	0.7586
ARM	0.3585	0.5016	0.6385	0.6987	0.7401
H&H	0.1978	0.4563	0.6188	0.6918	0.7275
L&L	0.3120	0.4600	0.6026	0.6741	0.7183
True	0.3618	0.5128	0.6435	0.7002	0.7348

Table 5.2.4.1.1 Overall estimates from 40-item test and all shapes: DA.

Method	Test Length				
	26	46	0	64	62
RM	0.8280	0.8298	0.8268	0.8280	0.8256
ARM	0.8087	0.8120	0.8116	0.8118	0.8103
H&H	0.8068	0.8068	0.8102	0.7998	0.8070
L&L	0.7950	0.7970	0.7966	0.8059	0.7782
True	0.8332	0.8338	0.8308	0.8280	0.8170

Table 5.2.4.1.2 Overall estimates from 40-item test and all shapes: DC.

Method	Test Length				
	26	46	0	64	62
RM	0.7572	0.7597	0.7557	0.7574	0.7542
ARM	0.7372	0.7395	0.7379	0.7390	0.7369
H&H	0.7286	0.7282	0.7322	0.7380	0.7274
L&L	0.7114	0.7162	0.7155	0.7281	0.6931
True	0.7474	0.7416	0.7466	0.7312	0.7220

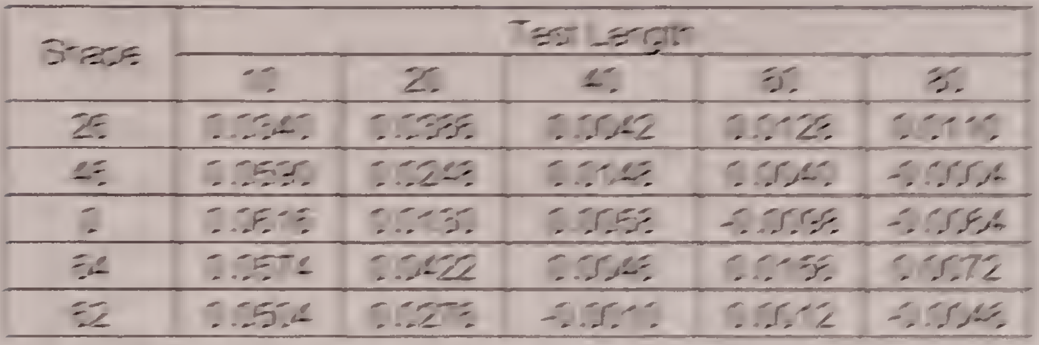
Table 5.2.4.1.3 Overall estimates from 40-item test and all shapes: Kappa.

Method	Test Length				
	26	46	0	64	62
RM	0.6608	0.6643	0.6609	0.6612	0.6567
ARM	0.6355	0.6398	0.6385	0.6388	0.6352
H&H	0.6161	0.6167	0.6188	0.6296	0.6096
L&L	0.5979	0.6065	0.6026	0.5966	0.5715
True	0.6749	0.6667	0.6661	0.6543	0.6238

Table 5.2.5.1.1 Differences between corrected and uncorrected H&H: DA

Shape	Test Length				
	10	20	40	60	80
26	0.0408	0.0252	0.0088	0.0058	0.0098
46	0.0480	0.0246	0.0012	0.0036	0.0050
0	0.0648	0.0258	0.0063	0.0014	-0.0020
64	0.0588	0.0238	0.0232	0.0114	0.0104
62	0.0554	0.0128	-0.0004	0.0074	0.0062

Table 5.2.5.1.2 Differences between corrected and uncorrected H&H: DC.



Shape	Test Length				
	10	20	40	60	80
25	0.0340	0.0386	0.0042	0.0128	0.0110
45	0.0590	0.0248	0.0148	0.0040	-0.0004
0	0.0616	0.0130	0.0058	-0.0098	-0.0064
64	0.0574	0.0422	0.0046	0.0168	0.0072
52	0.0504	0.0276	-0.0010	0.0012	-0.0046

Table 5.2.5.1.3 Differences between corrected and uncorrected H&H: Kappa.

Shape	Test Length				
	10	20	40	60	80
26	0.1336	0.0738	0.0109	0.0202	0.0170
46	0.1372	0.0558	0.0259	0.0093	0.0014
0	0.1759	0.0635	0.0200	0.0066	0.0051
64	0.1336	0.0784	0.0134	0.0255	0.0122
62	0.1506	0.0631	0.0075	0.0071	-0.0036

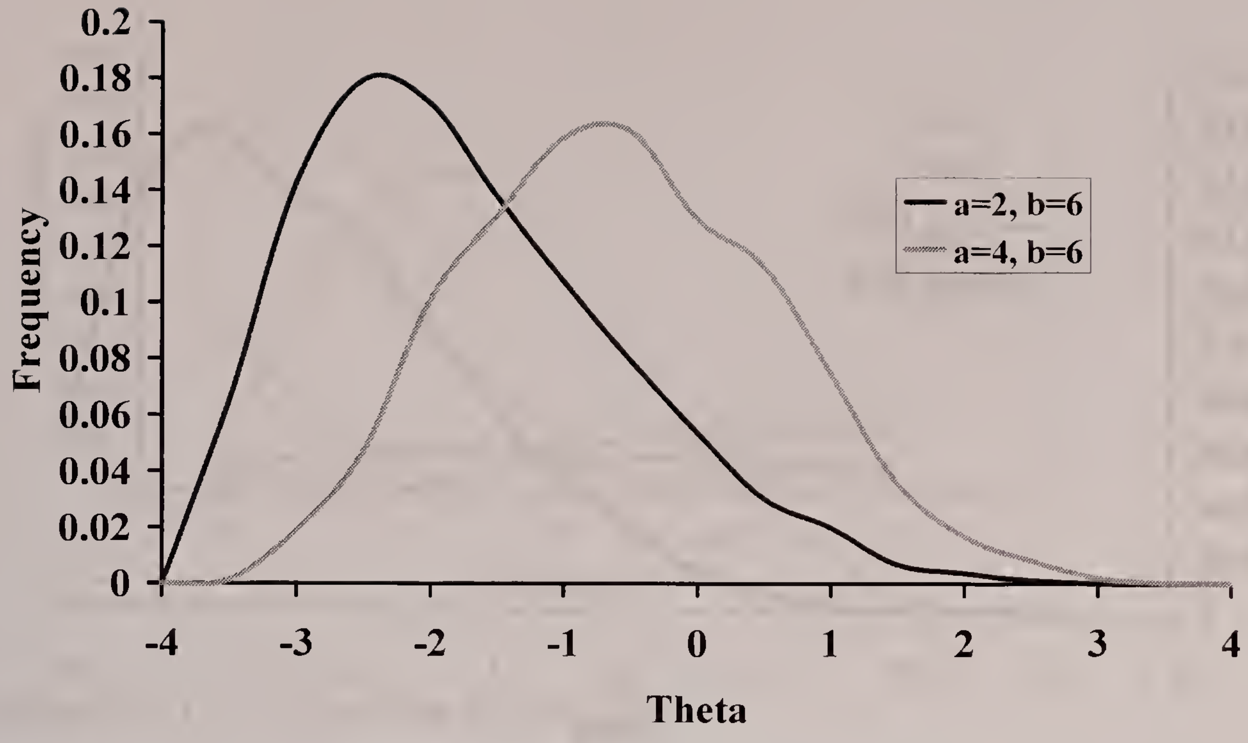


Figure 5.1.2.1 Two negatively skewed theta distributions.

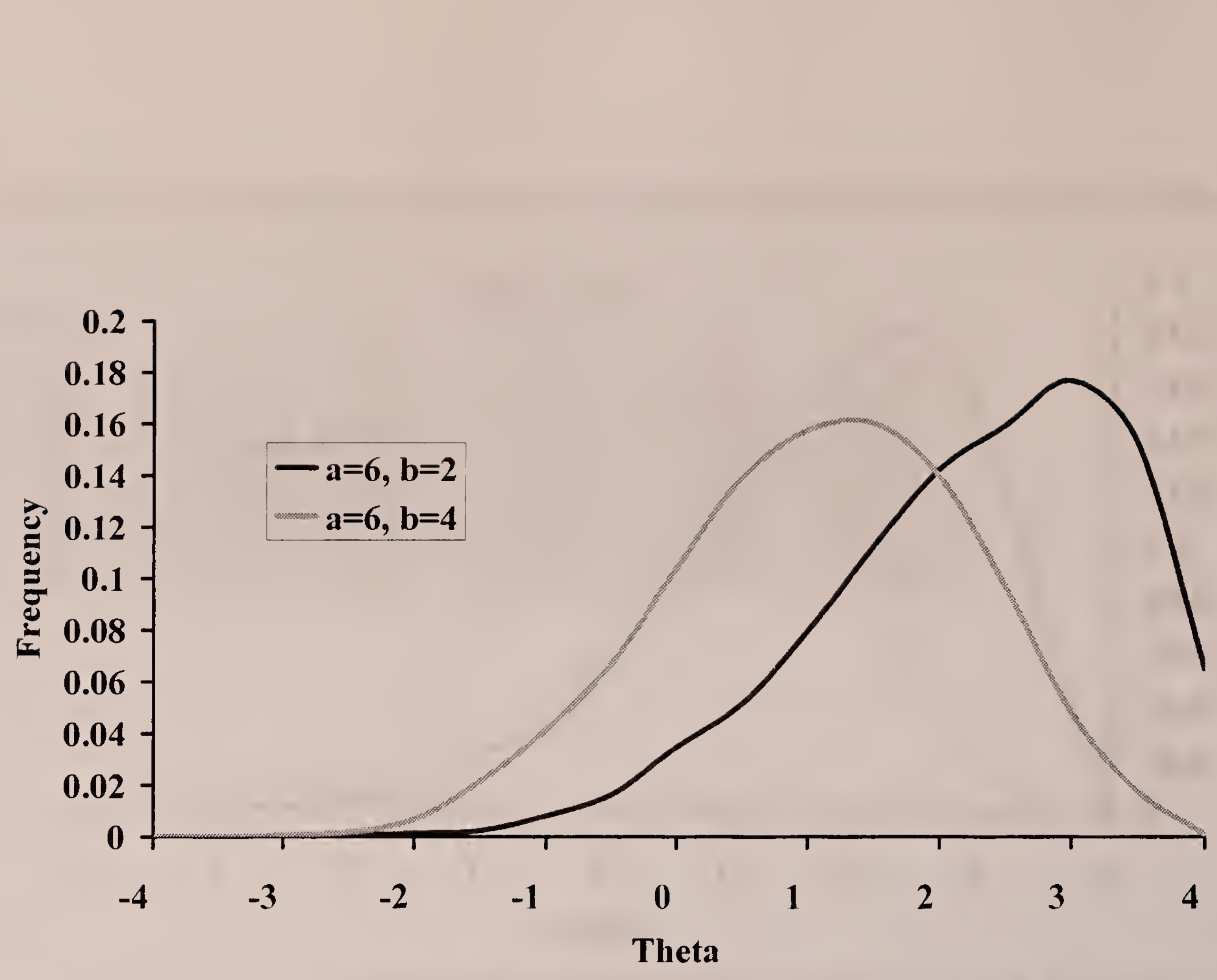


Figure 5.1.2.2 Two positively skewed theta distributions.

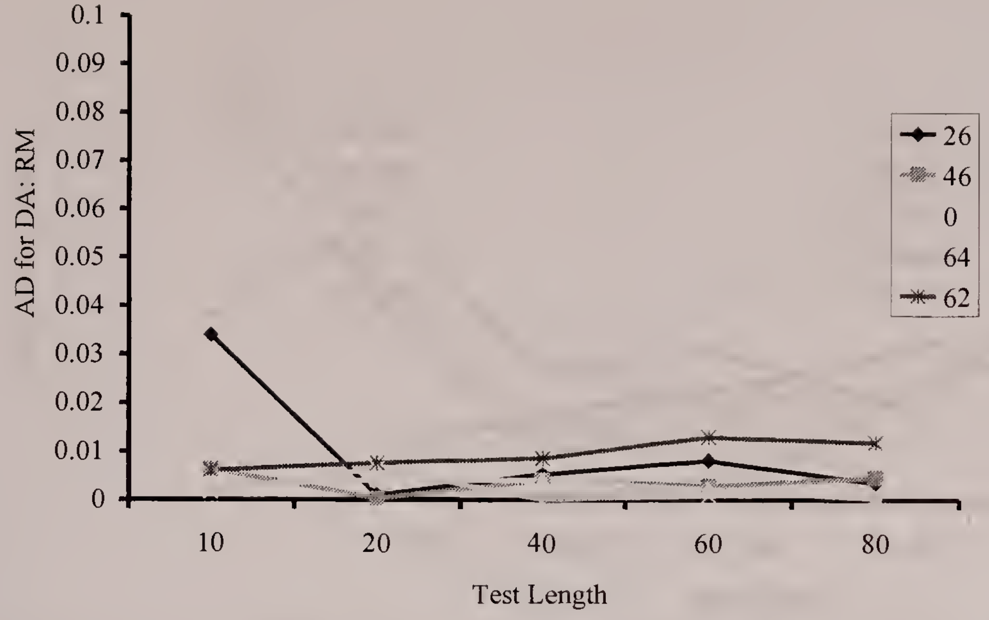


Figure 5.2.1.1.1 AD values for DA: RM.

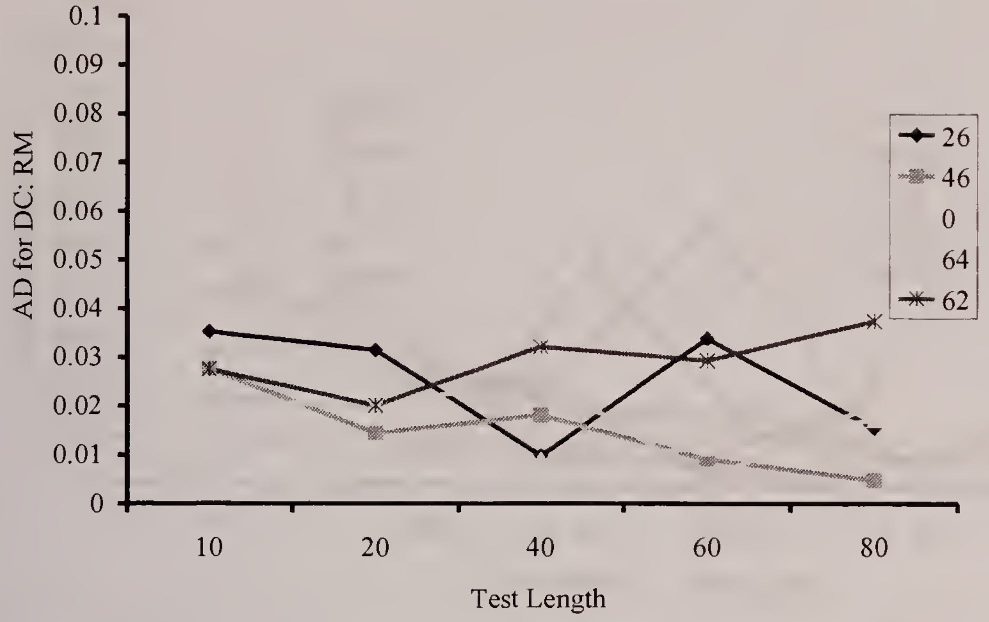


Figure 5.2.1.1.2 AD values for DC: RM.

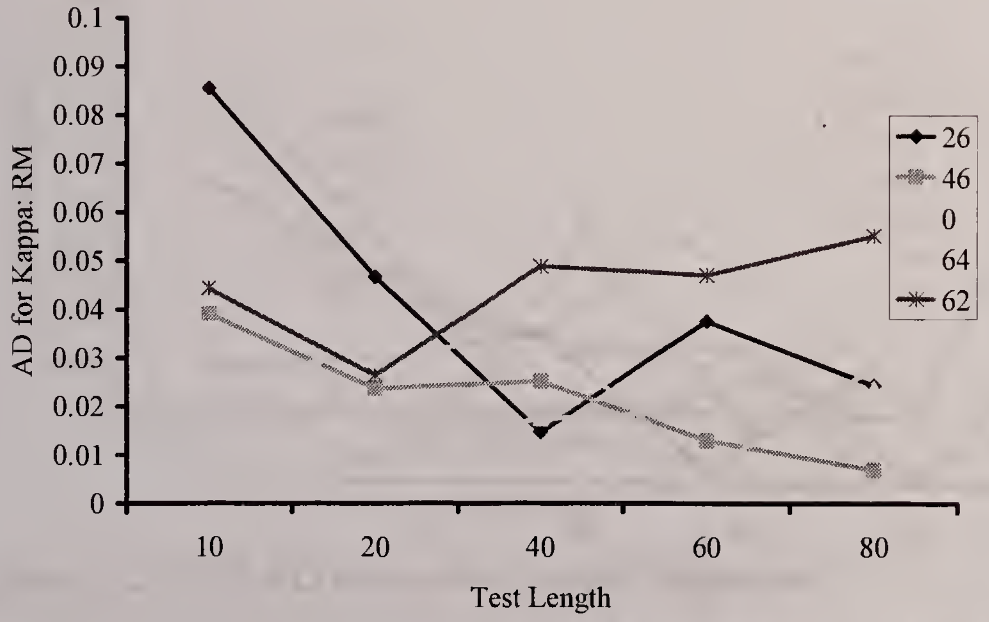


Figure 5.2.1.1.3 AD values for Kappa: RM.

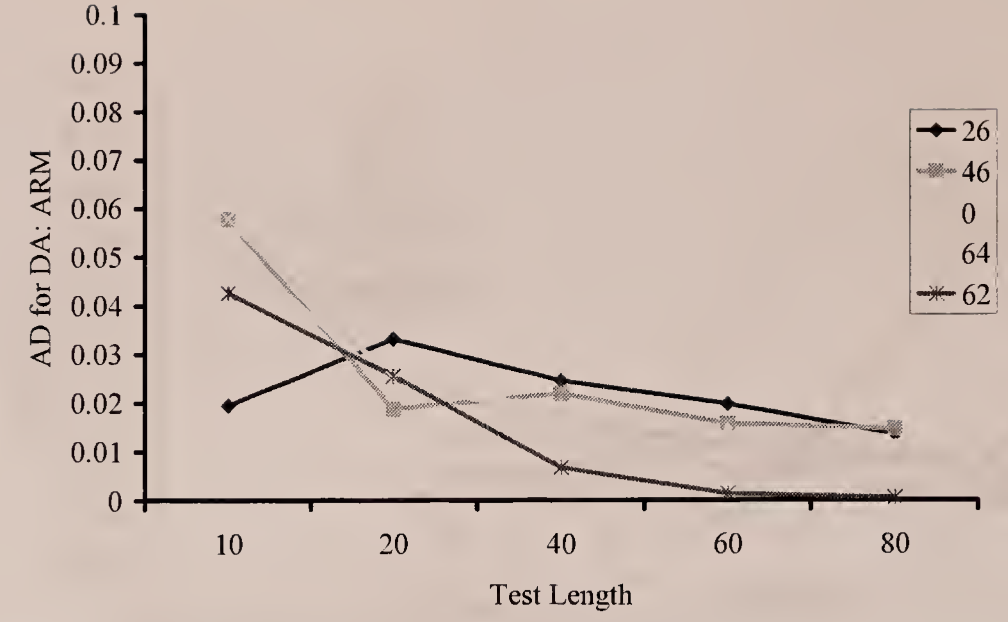


Figure 5.2.1.2.1 AD values for DA: ARM.

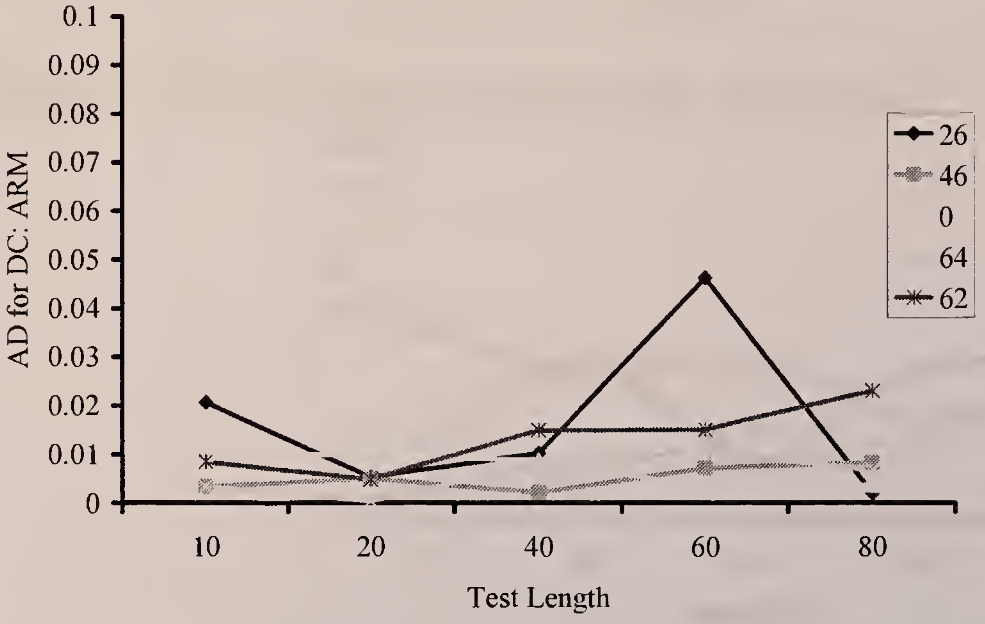


Figure 5.2.1.2.2 AD values for DC: ARM.

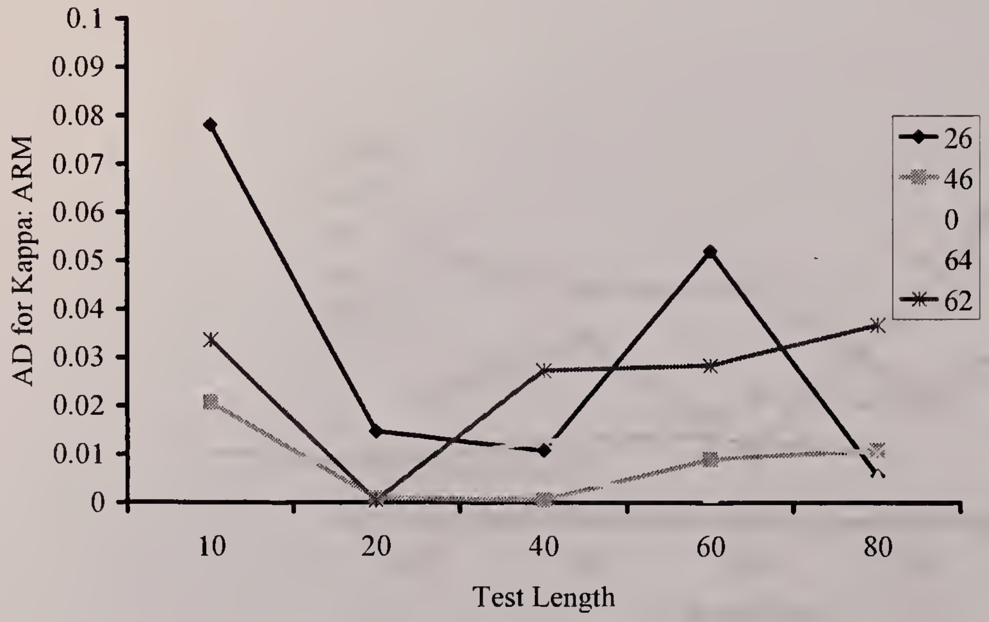


Figure 5.2.1.2.3 AD values for Kappa: ARM.

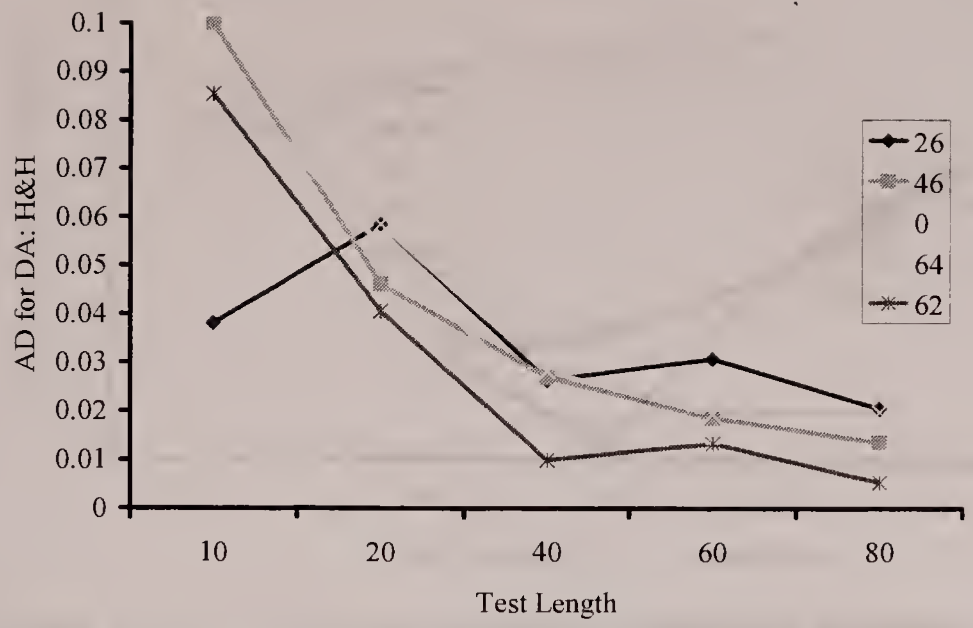


Figure 5.2.1.3.1 AD values for DA: H&H.

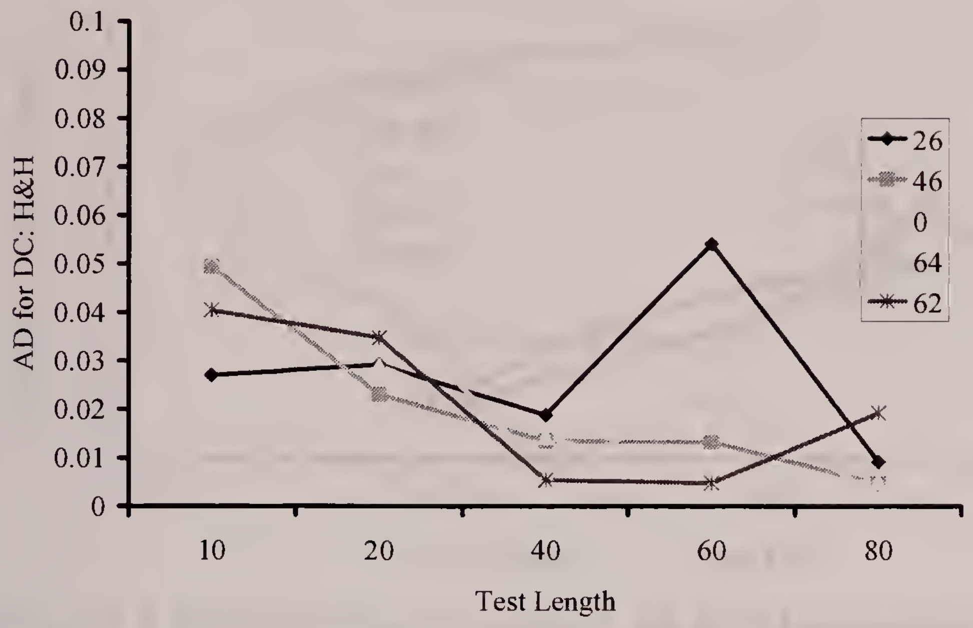


Figure 5.2.1.3.2 AD values for DC: H&H.

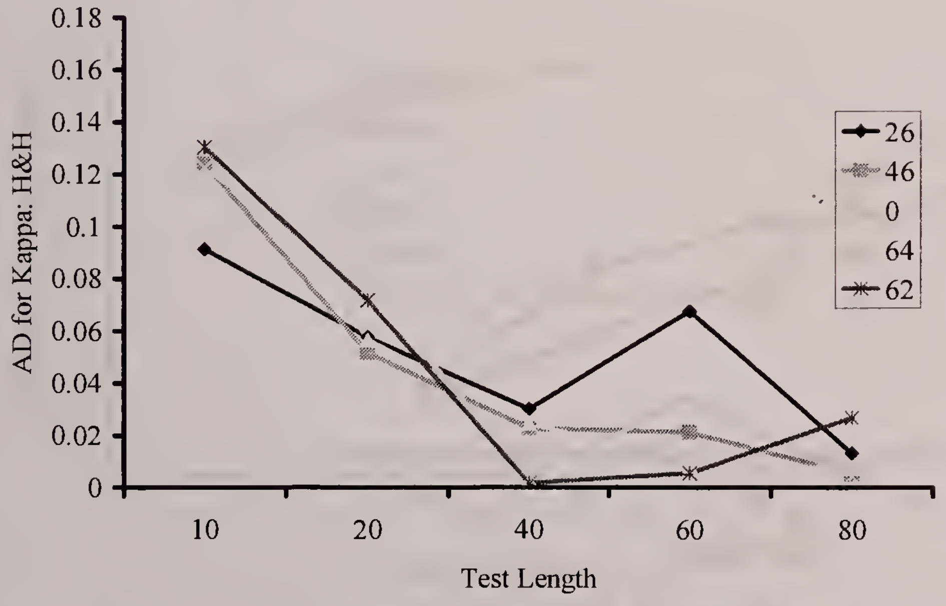


Figure 5.2.1.3.3 AD values for Kappa: H&H.

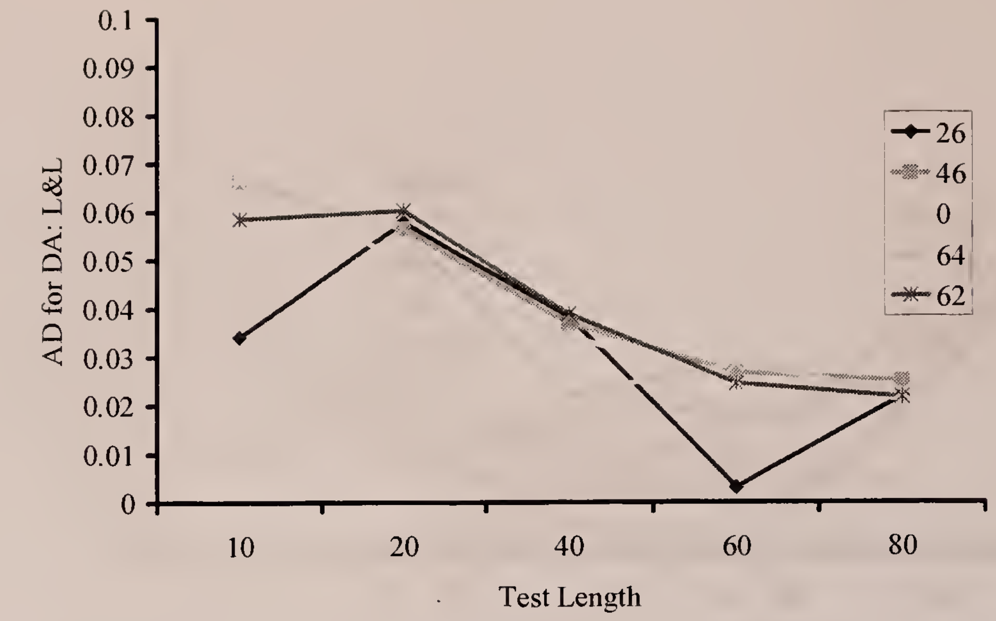


Figure 5.2.1.4.1 AD values for DA: L&L.

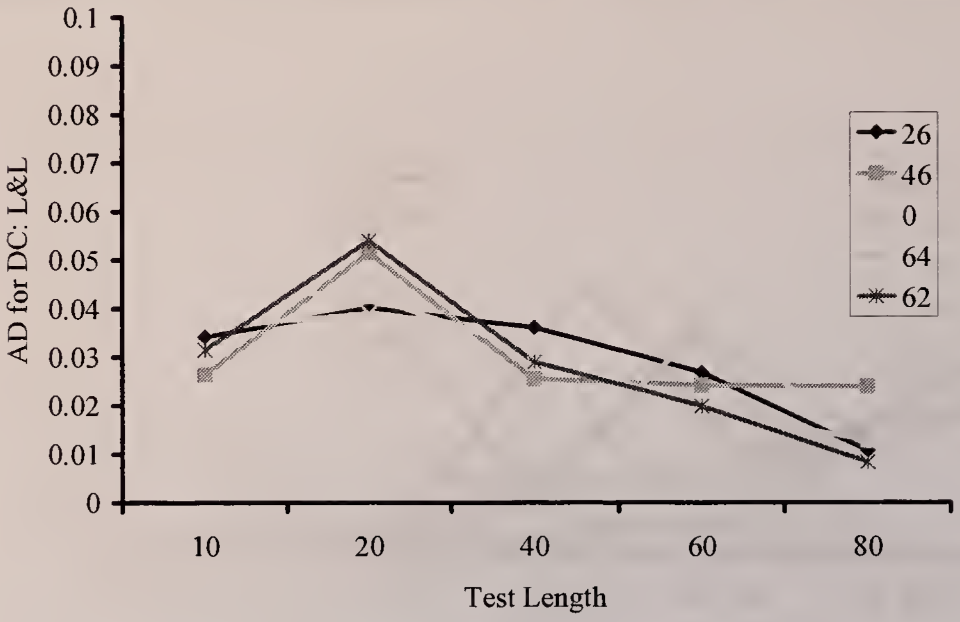


Figure 5.2.1.4.2 AD values for DC: L&L.

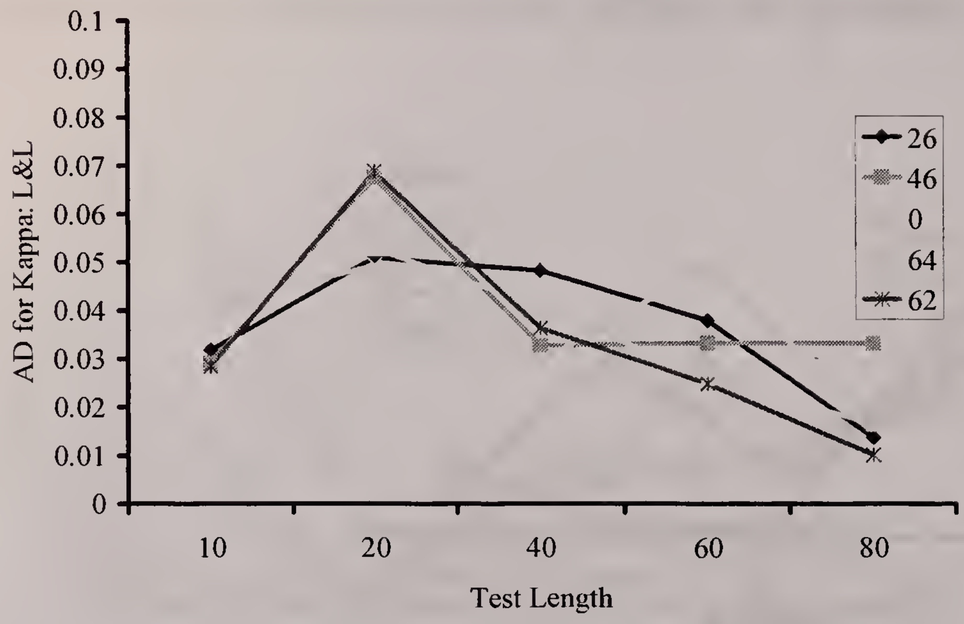


Figure 5.2.1.4.3 AD values for Kappa: L&L.

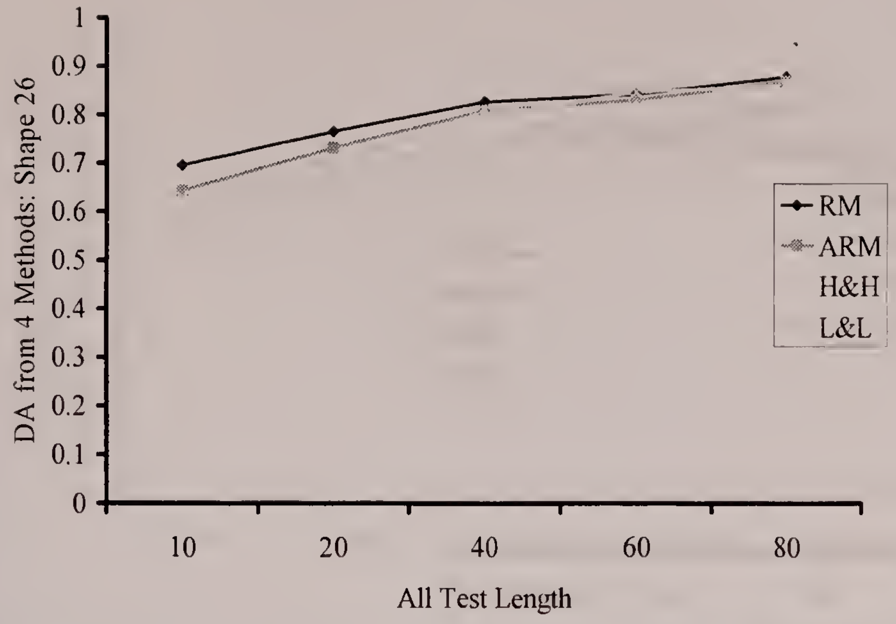


Figure 5.2.2.1.1 DA from 4 methods: positively skewed.

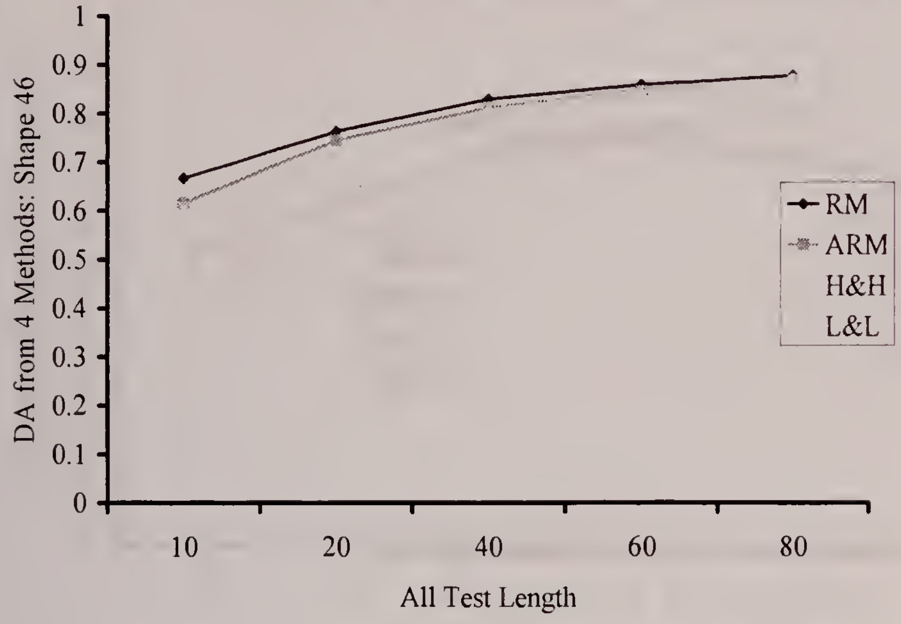


Figure 5.2.2.1.2 DA from 4 methods: slightly positively skewed.

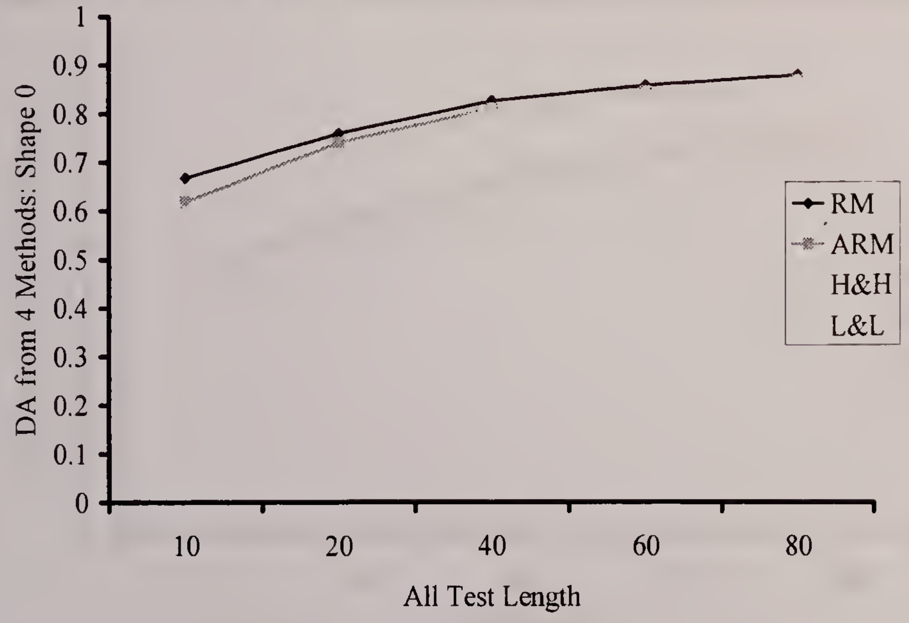


Figure 5.2.2.1.3 DA from 4 methods: normal.

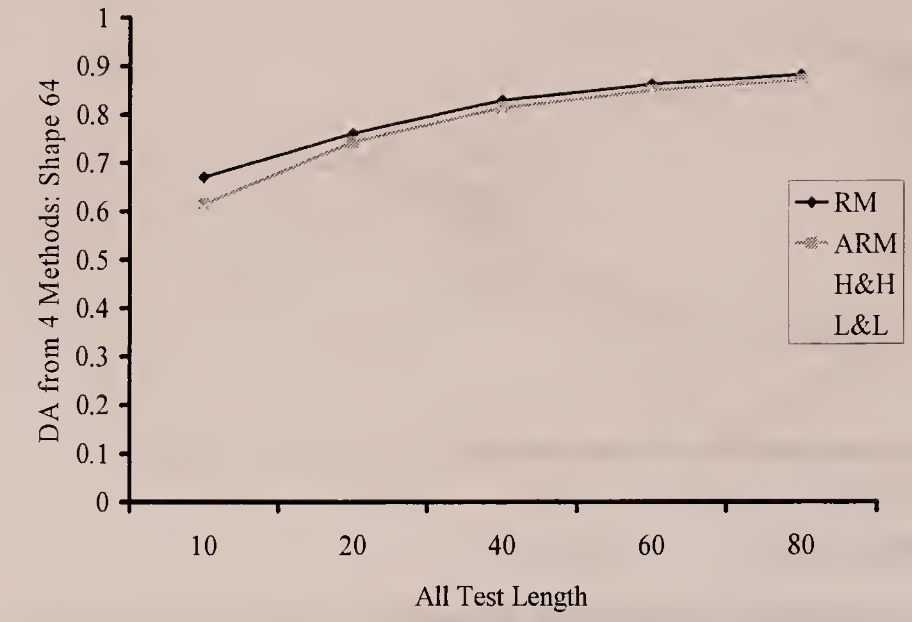


Figure 5.2.2.1.4 DA from 4 methods: slightly negatively skewed.

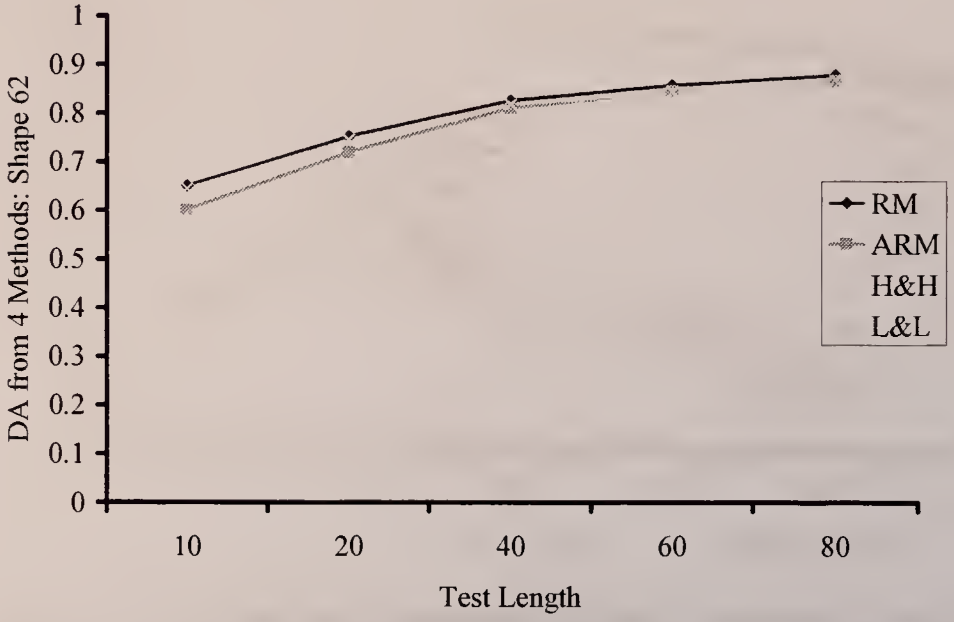


Figure 5.2.2.1.5 DA from 4 methods: negatively skewed.

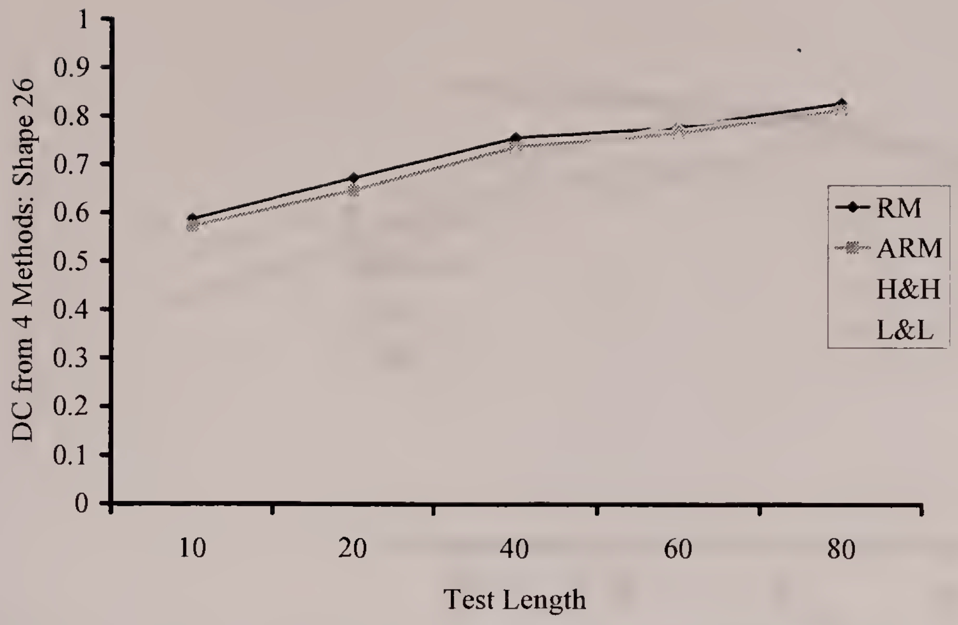


Figure 5.2.2.2.1 DC from 4 methods: positively skewed.

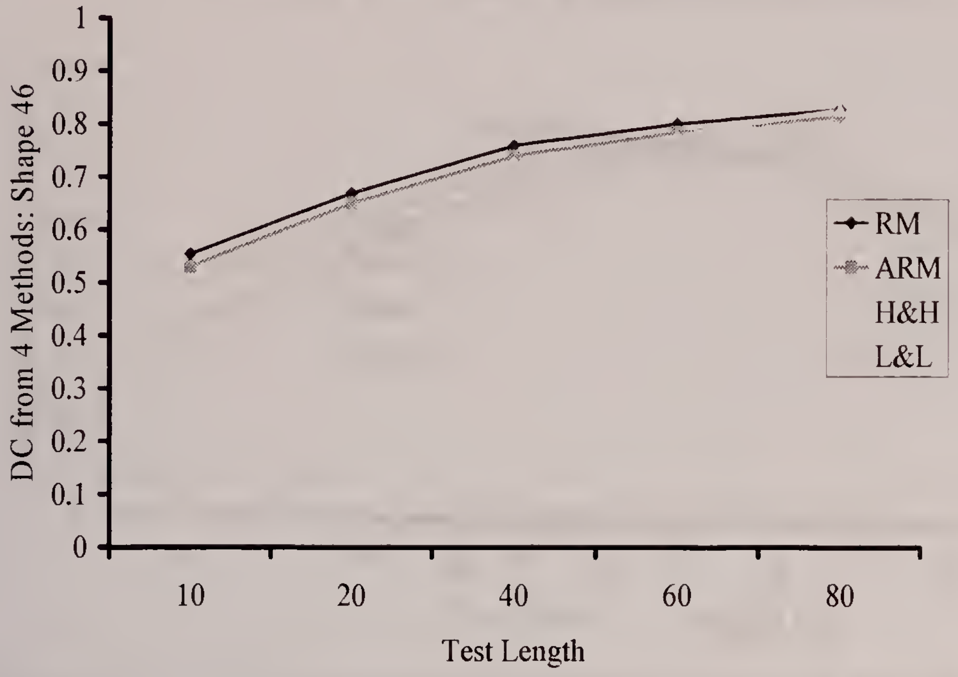


Figure 5.2.2.2.2 DC from 4 methods: slightly positively skewed.

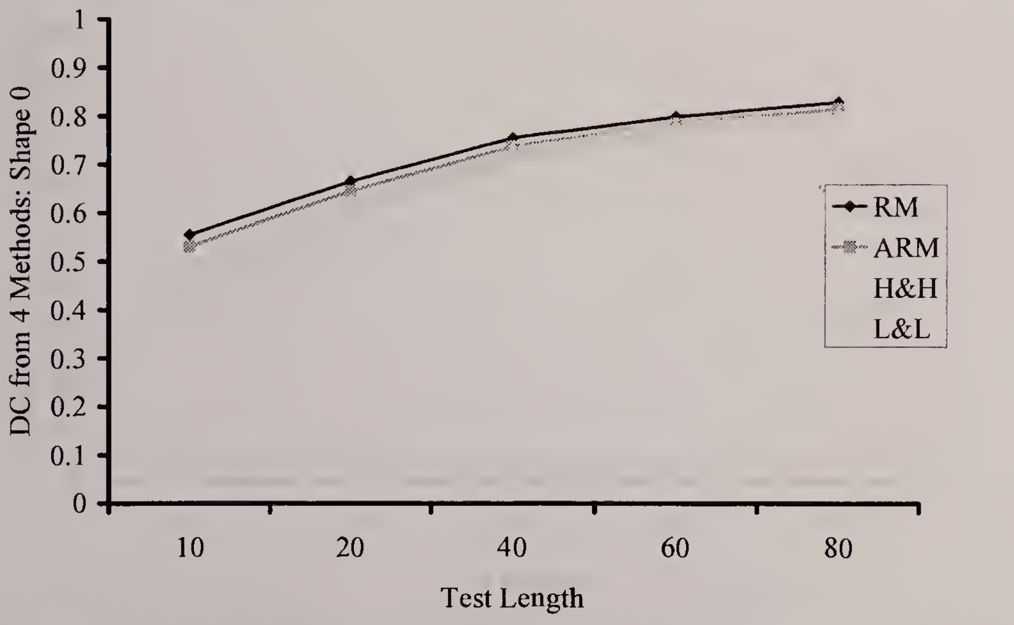


Figure 5.2.2.2.3 DC from 4 methods: normal.

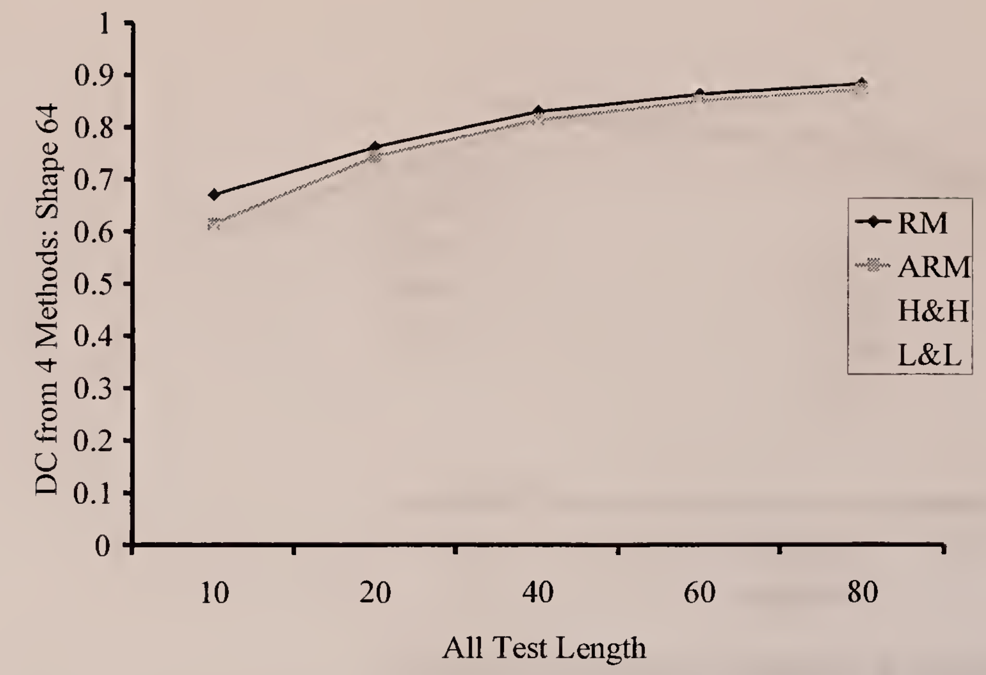


Figure 5.2.2.2.4 DC from 4 methods: slightly negatively skewed.

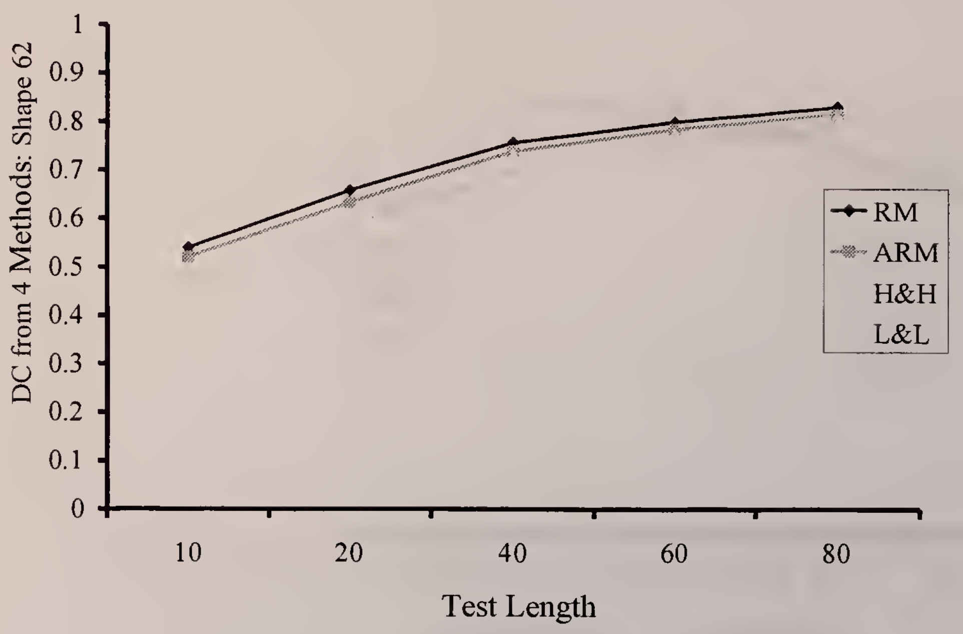


Figure 5.2.2.2.5 DC from 4 methods: negatively skewed.

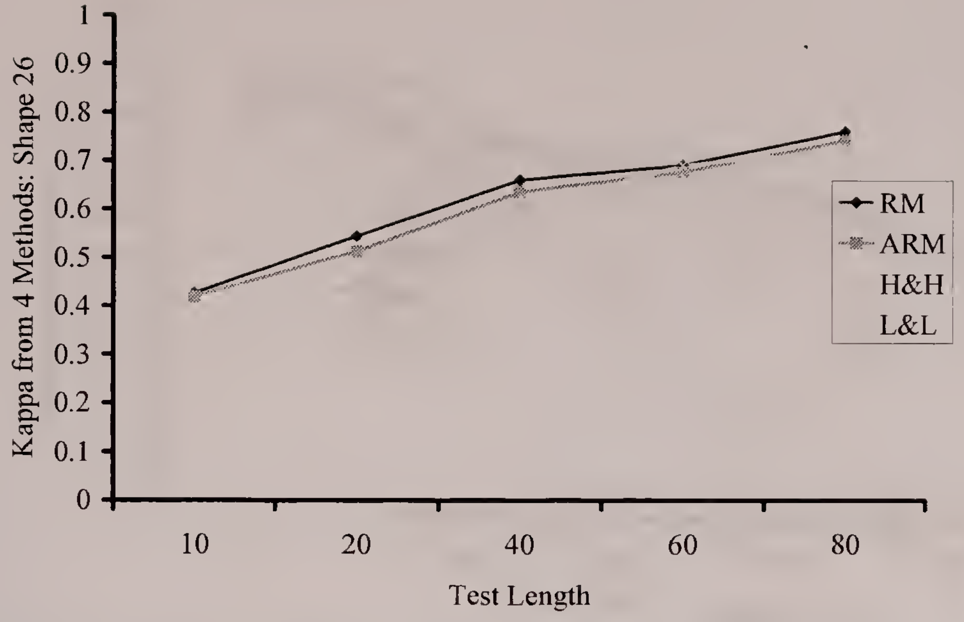


Figure 5.2.2.3.1 Kappa from 4 methods: positively skewed.

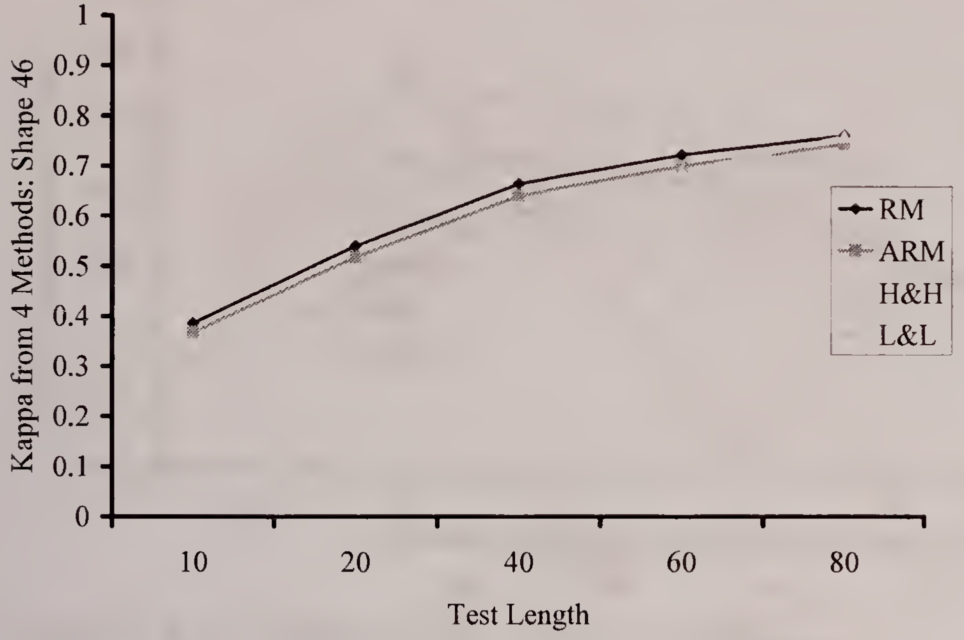


Figure 5.2.2.3.2 Kappa from 4 methods: slightly positively skewed.

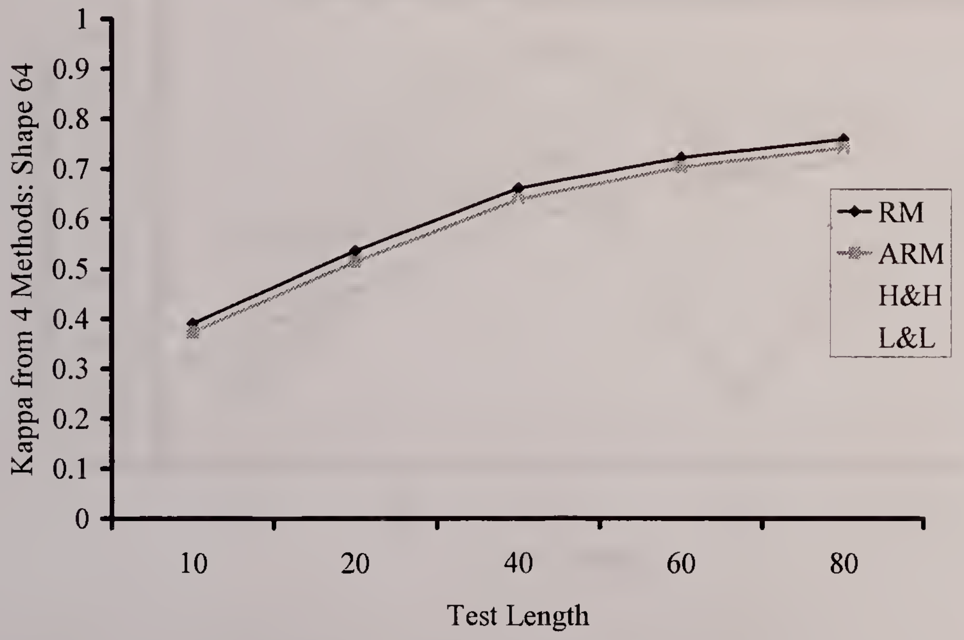


Figure 5.2.2.3.3 Kappa from 4 methods: slightly negatively skewed

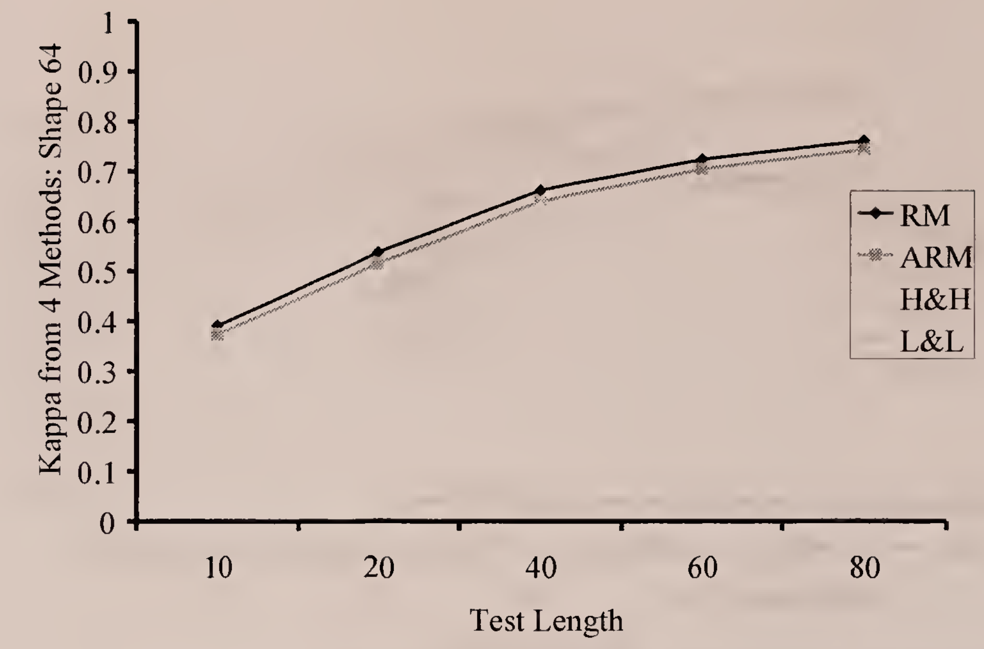


Figure 5.2.2.3.4 Kappa from 4 methods: slightly negatively skewed.

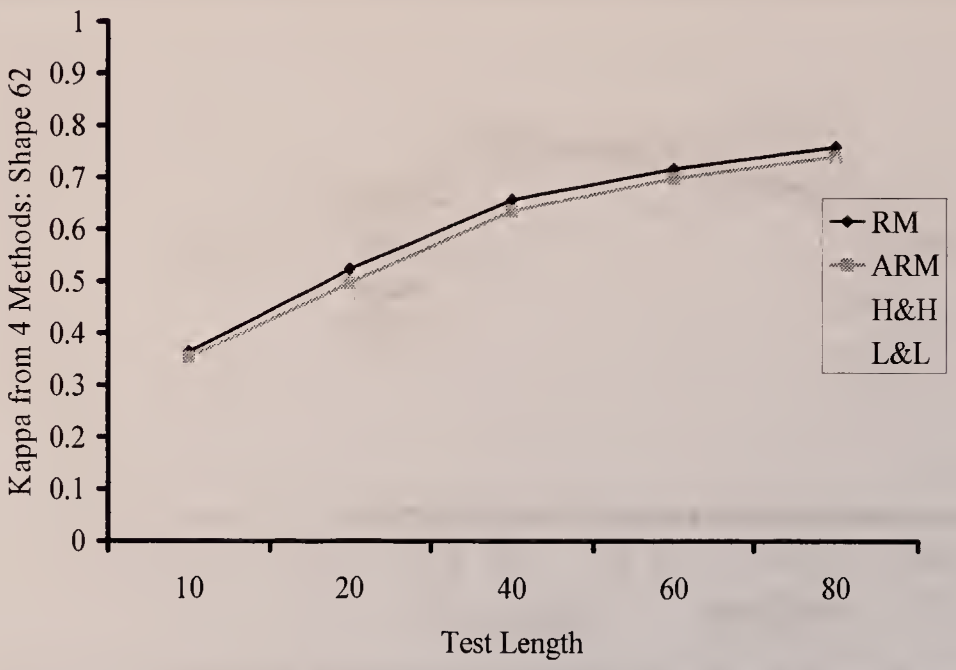


Figure 5.2.2.3.5 Kappa from 4 methods: negatively skewed.

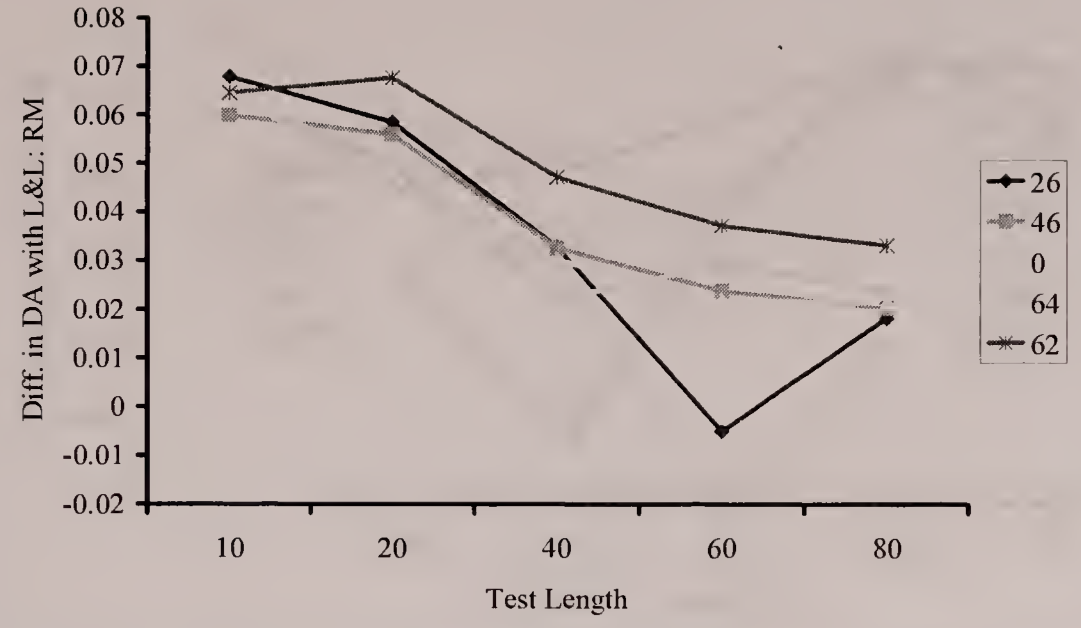


Figure 5.2.2.4.1 Diff. in DA: L&L vs. RM.

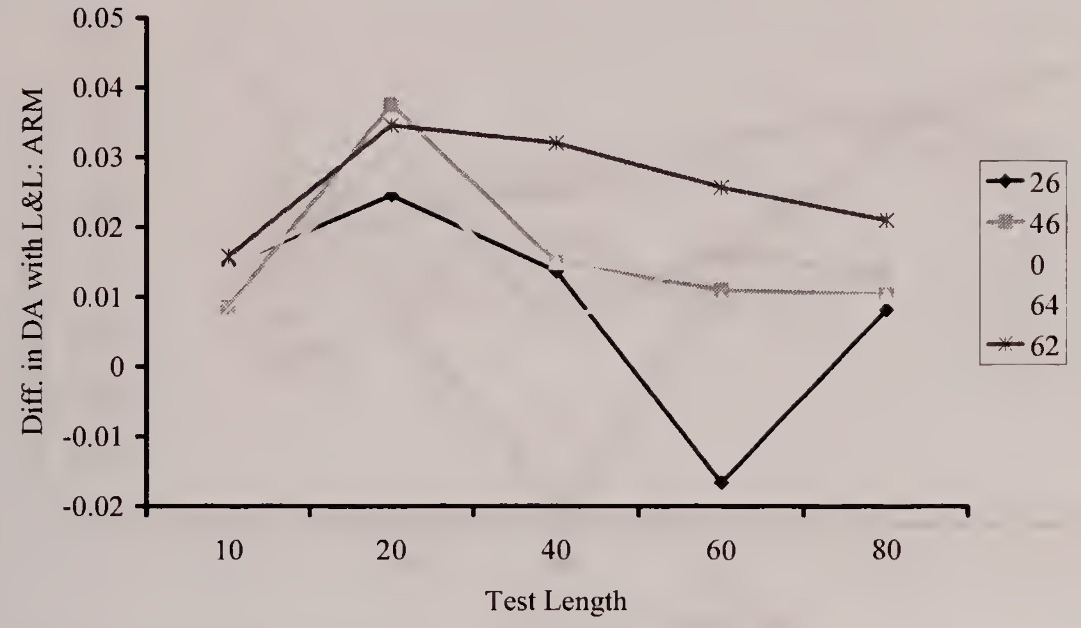


Figure 5.2.2.4.2 Diff. in DA: L&L vs. ARM.

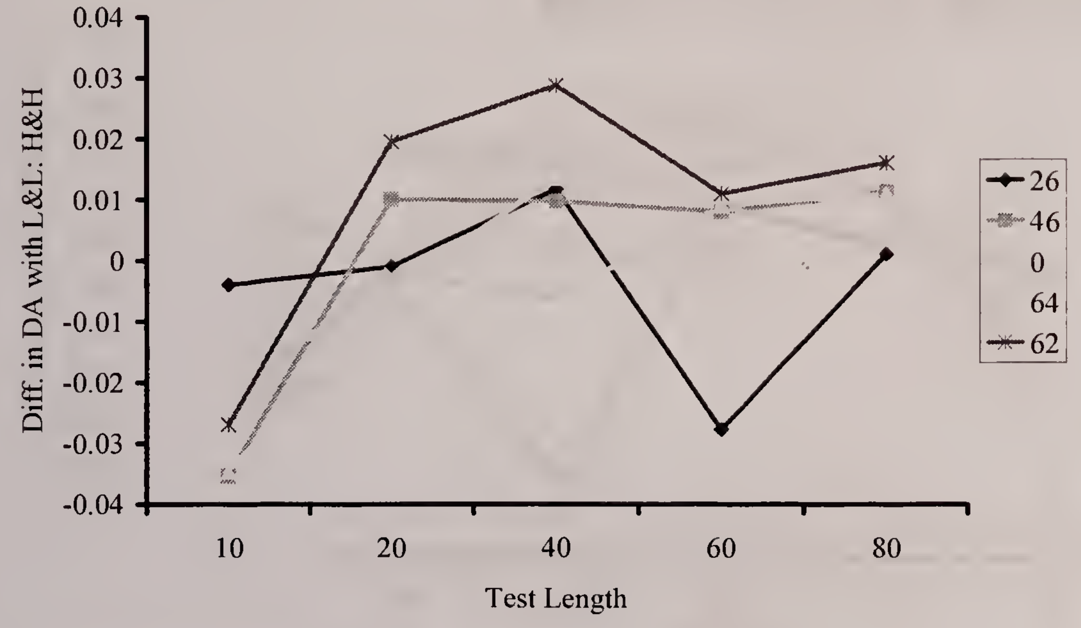


Figure 5.2.2.4.3 Diff. in DA: L&L vs. H&H.

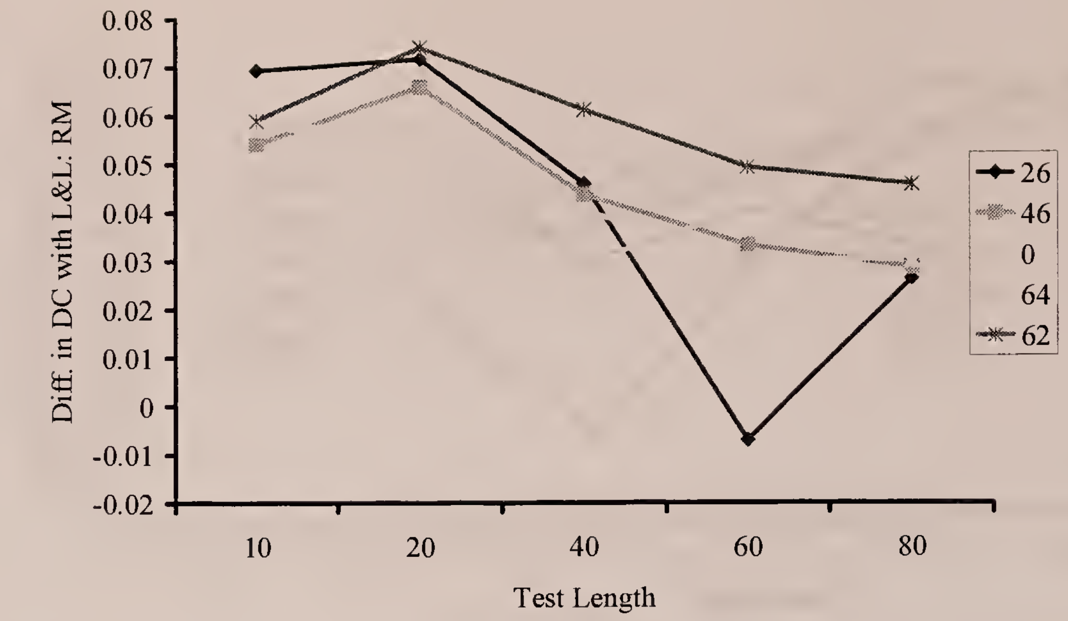


Figure 5.2.2.5.1 Diff. in DC: L&L vs. RM.

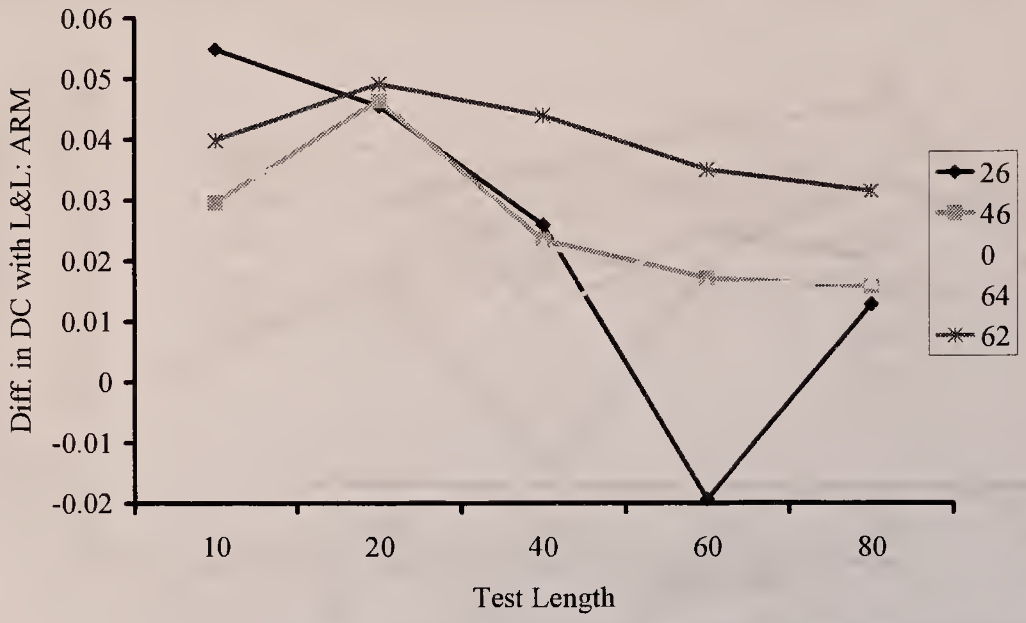


Figure 5.2.2.5.2 Diff. in DC: L&L vs. ARM.

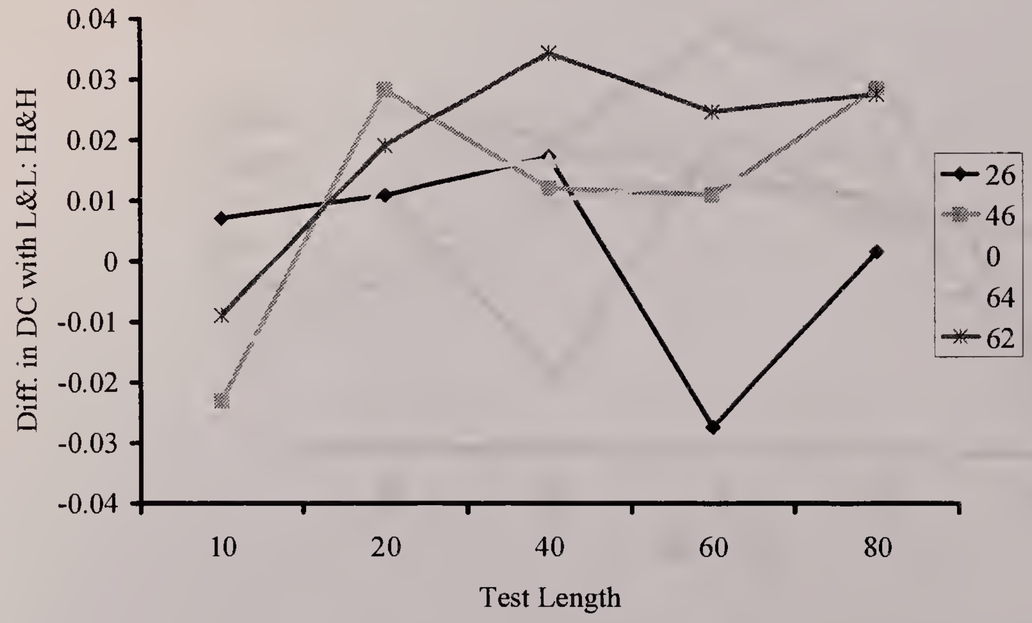


Figure 5.2.2.5.3 Diff. in DC: L&L vs. H&H.

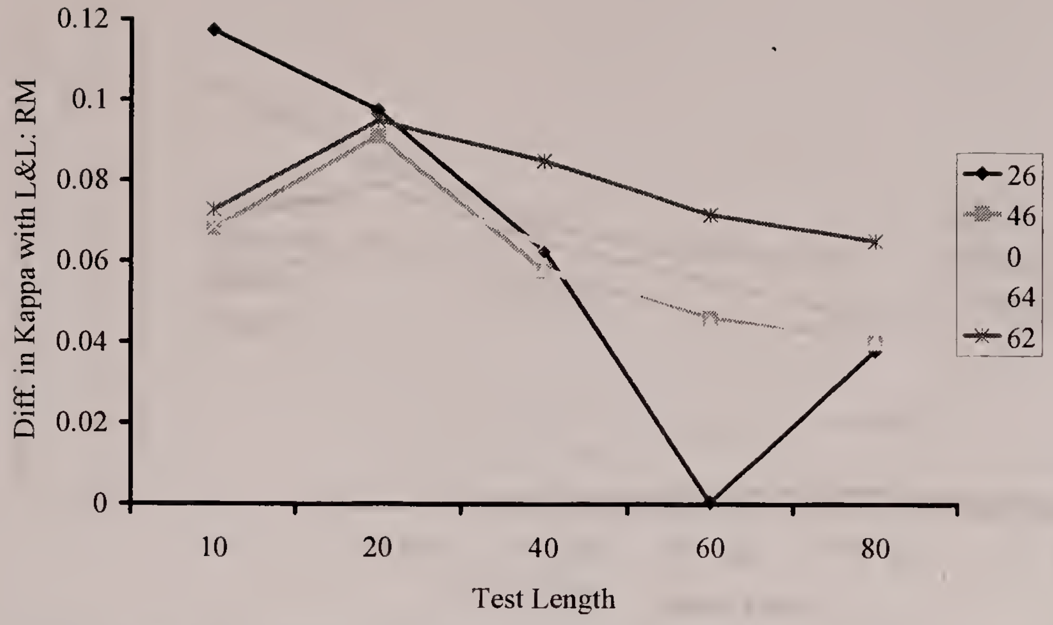


Figure 5.2.2.6.1 Diff. in Kappa: L&L vs. RM.

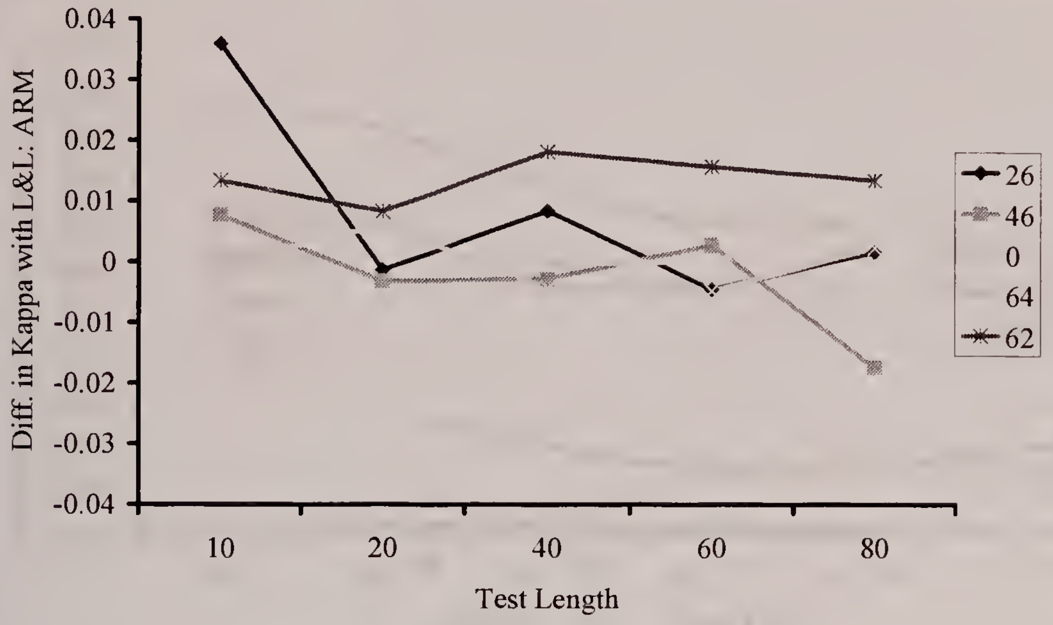


Figure 5.2.2.6.2 Diff. in Kappa: L&L vs. ARM.

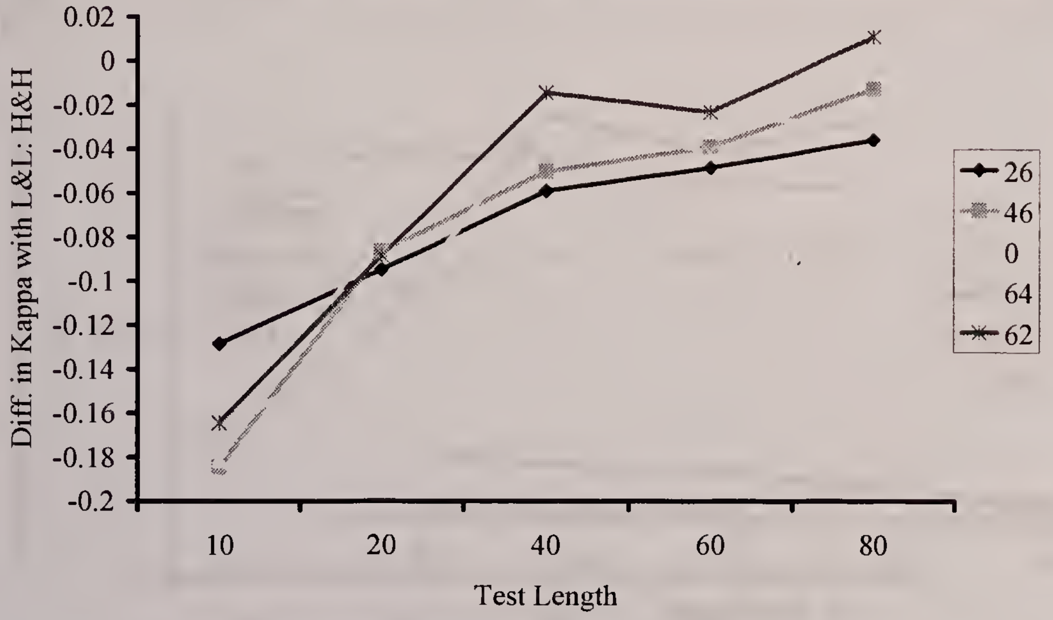


Figure 5.2.2.6.3 Diff. in Kappa: L&L vs. H&H.

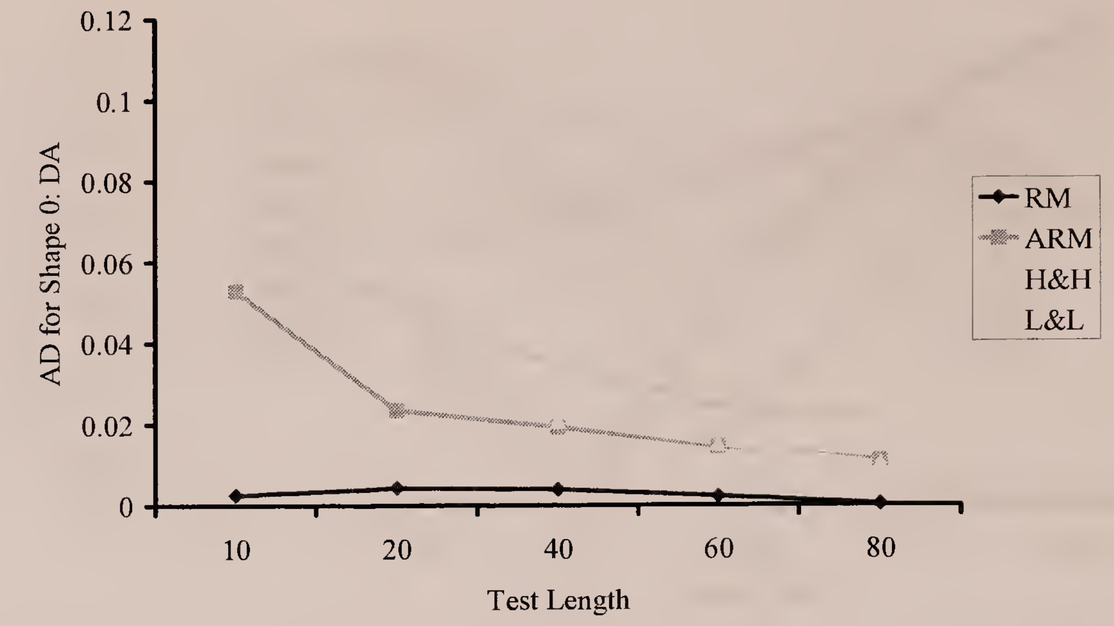


Figure 5.2.3.1.1 AD for the normal true score distribution shape: DA.

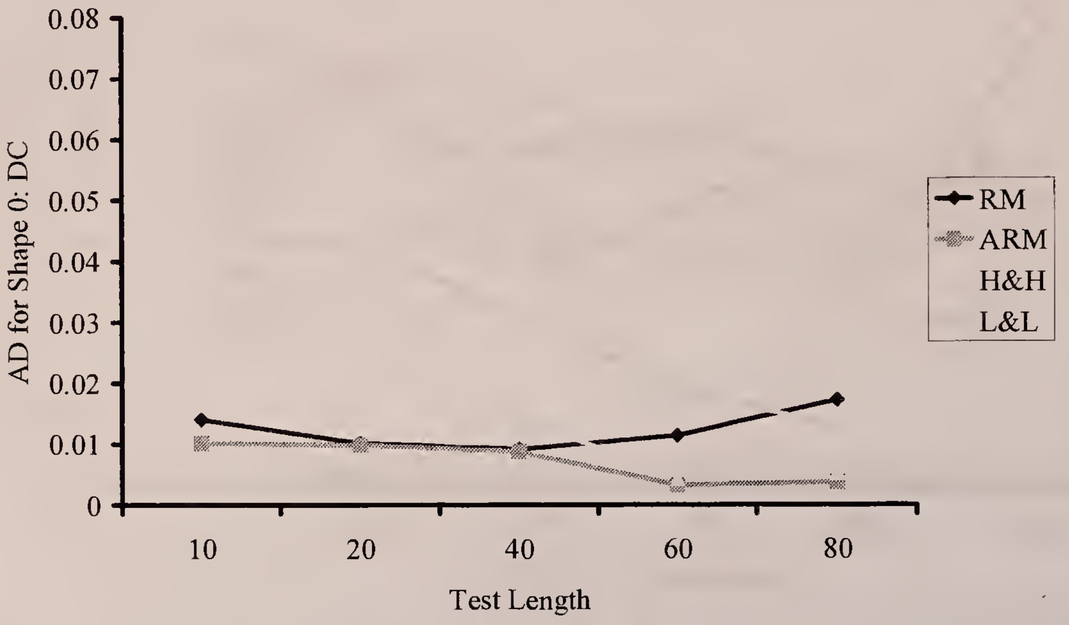


Figure 5.2.3.1.2 AD for the normal true score distribution shape: DC.

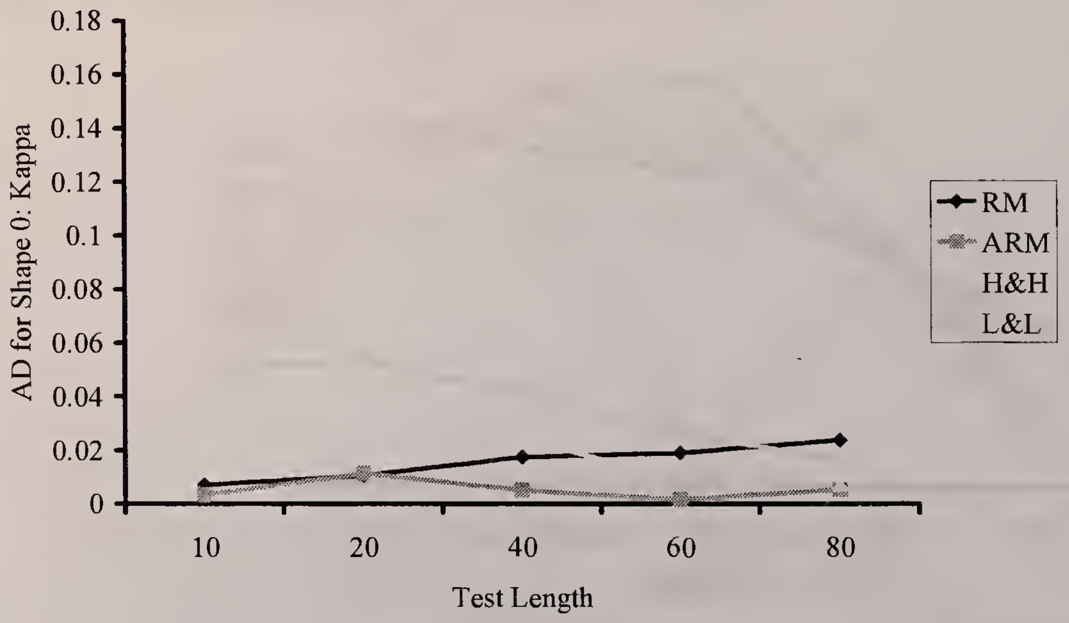


Figure 5.2.3.1.3 AD for the normal true score distribution shape: Kappa.

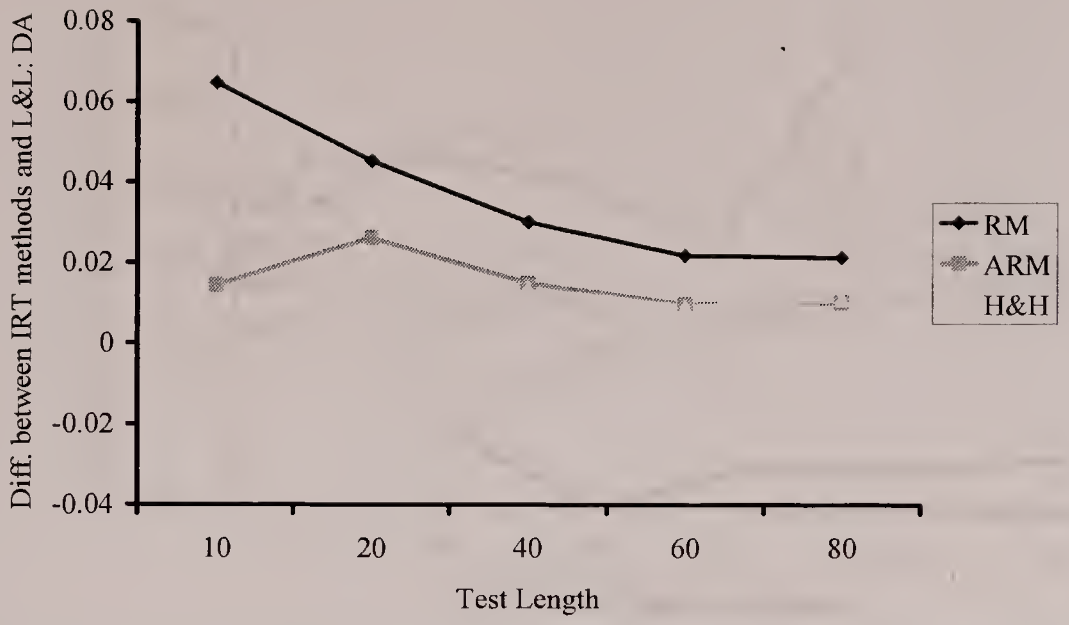


Figure 5.2.3.2.1 Diff. with L&L for the normal true score distribution shape: DA.

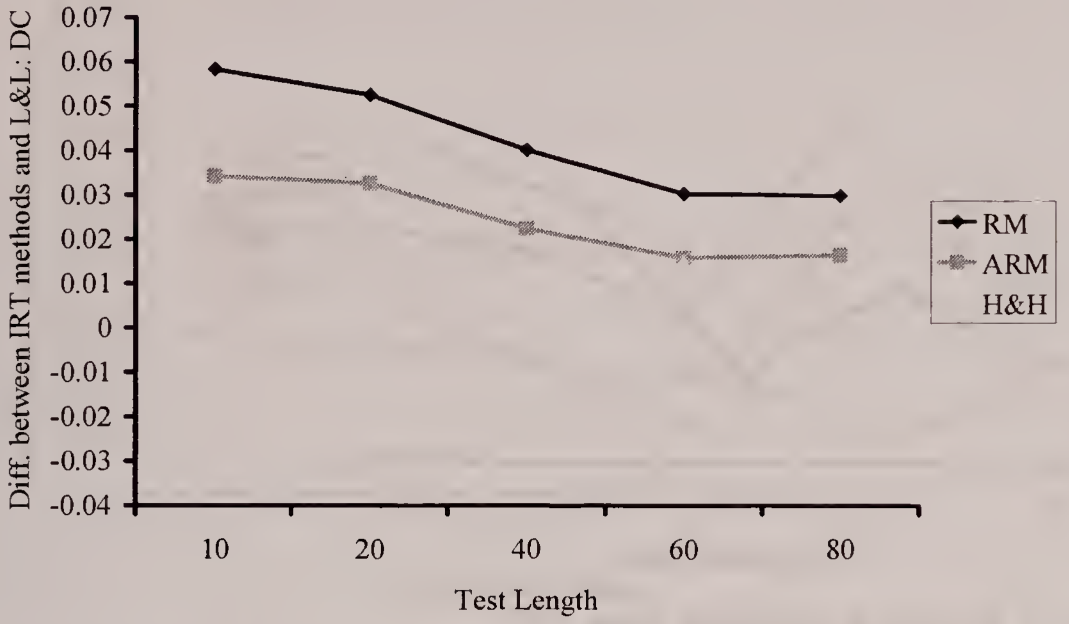


Figure 5.2.3.2.2 Diff. with L&L for the normal true score distribution shape: DC.

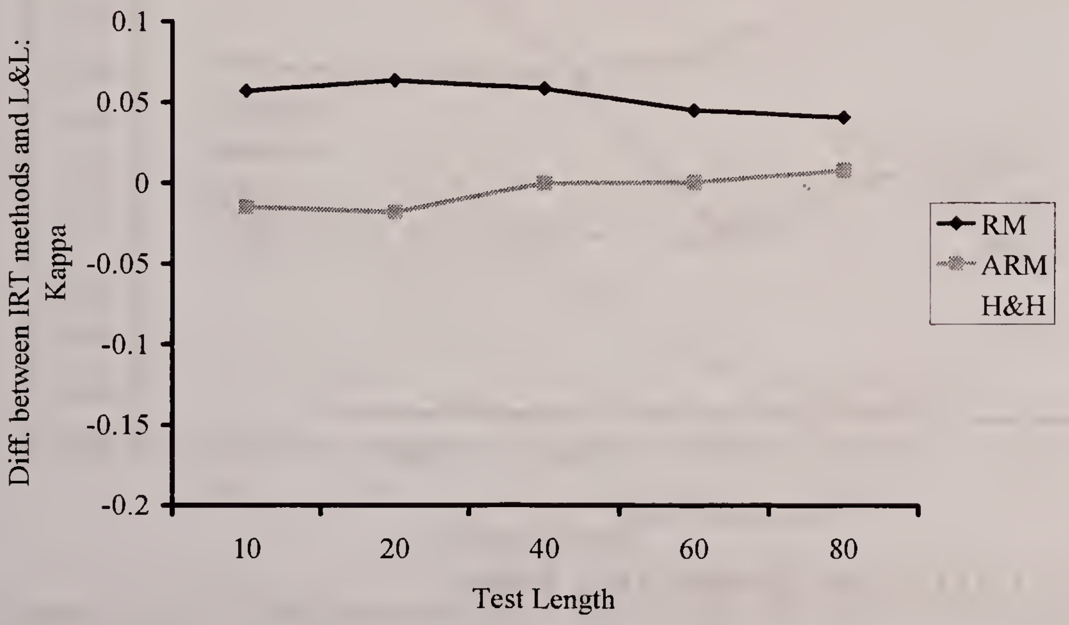


Figure 5.2.3.2.3 Diff. with L&L for the normal true score distribution shape: Kappa.

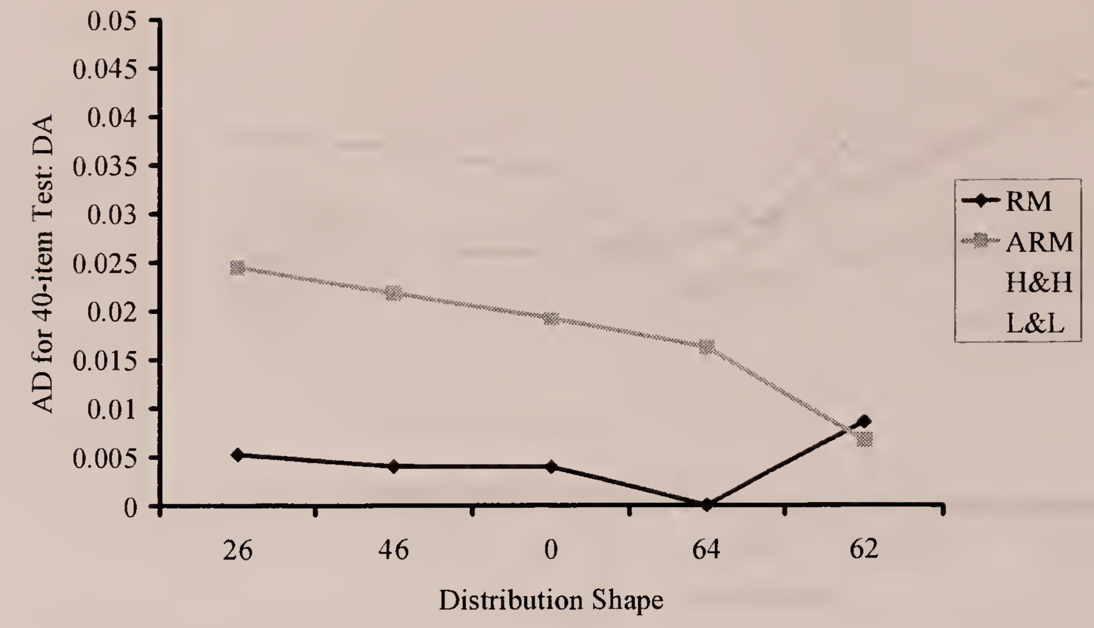


Figure 5.2.4.1.1 AD for the 40-item Test: DA.

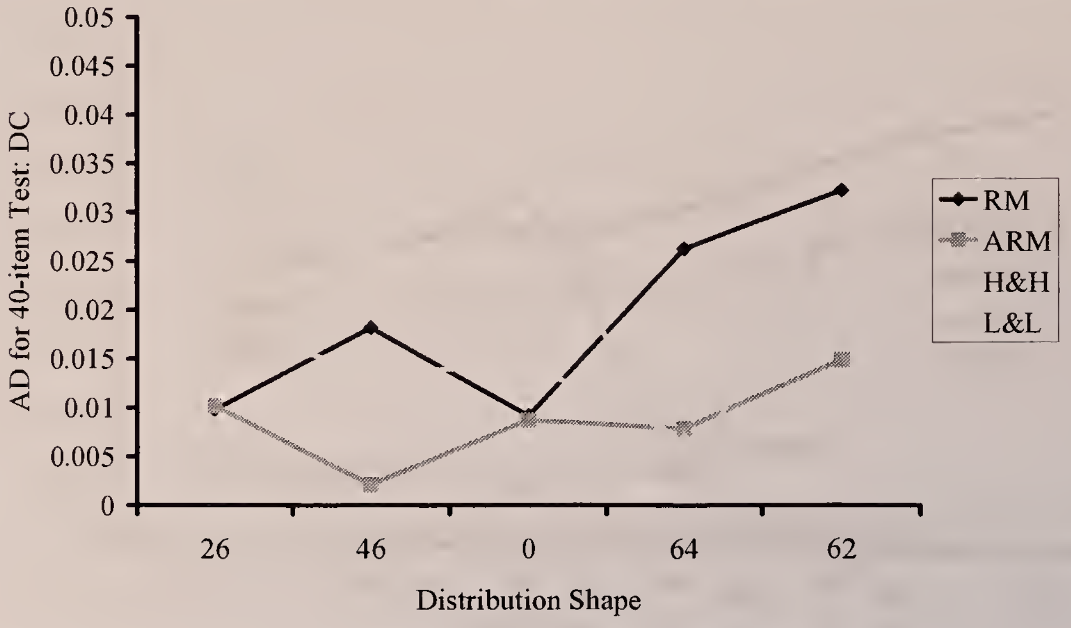


Figure 5.2.4.1.2 AD for the 40-item Test: DC.

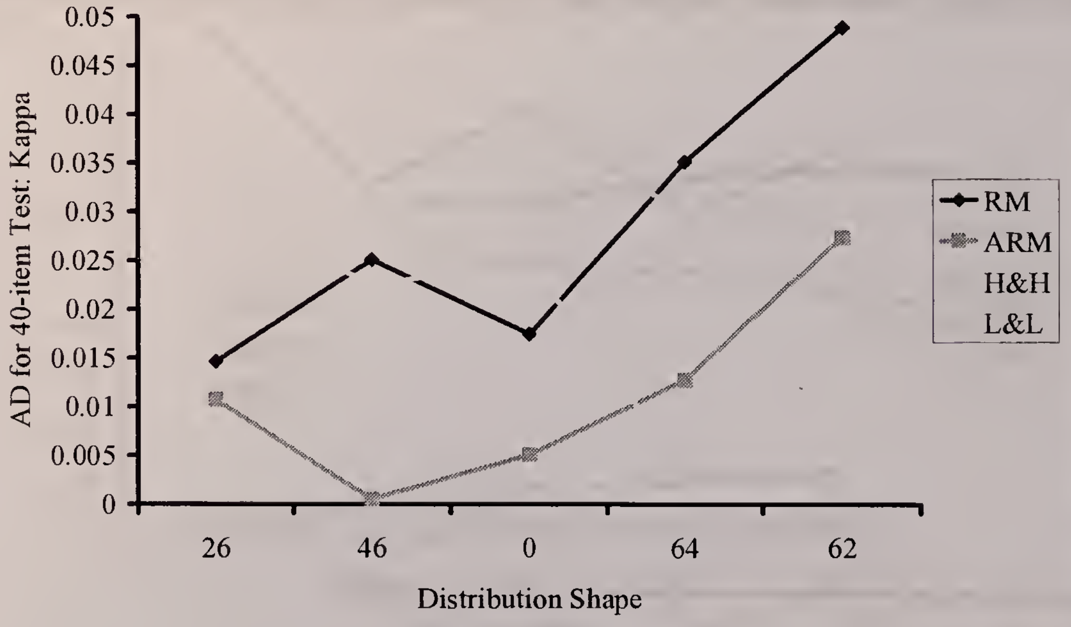


Figure 5.2.4.1.3 AD for the 40-item Test: Kappa.

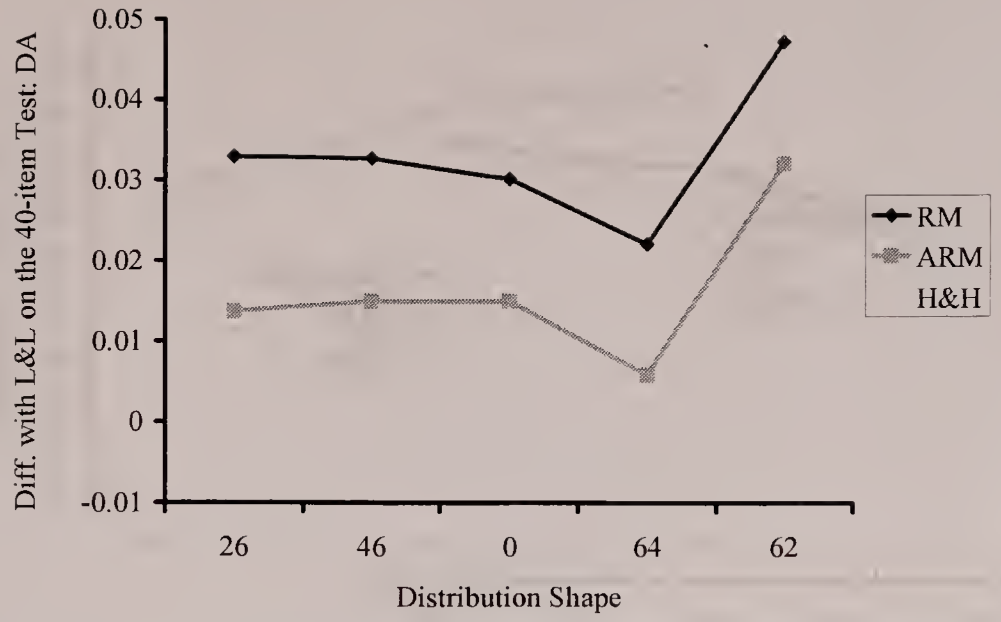


Figure 5.2.4.2.1 Difference with L&L for the 40-item Test: DA.

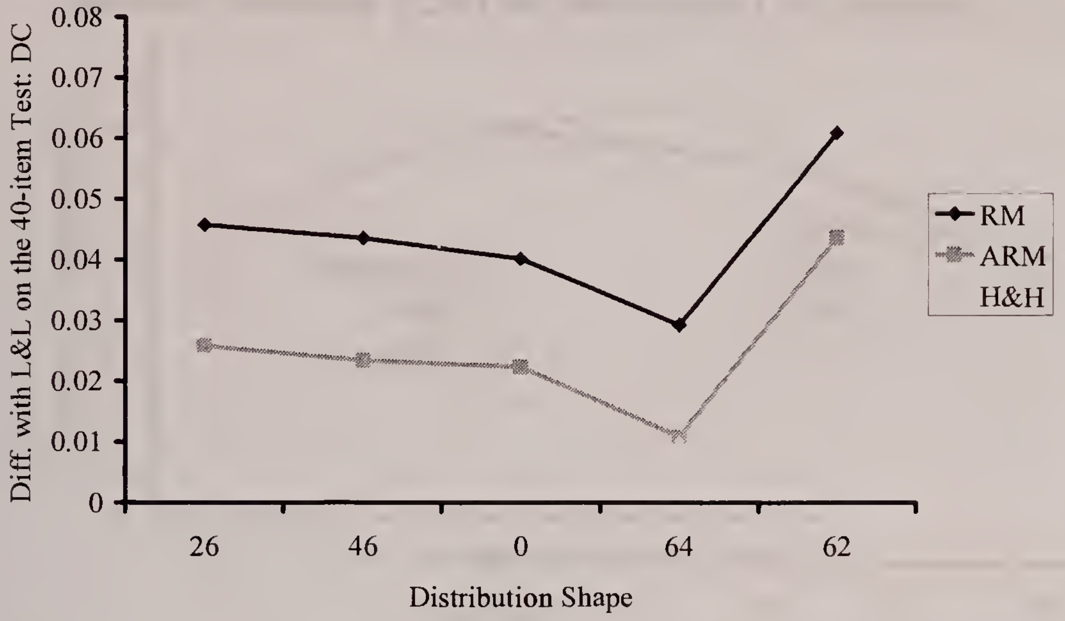


Figure 5.2.4.2.2 Difference with L&L for the 40-item Test: DC.

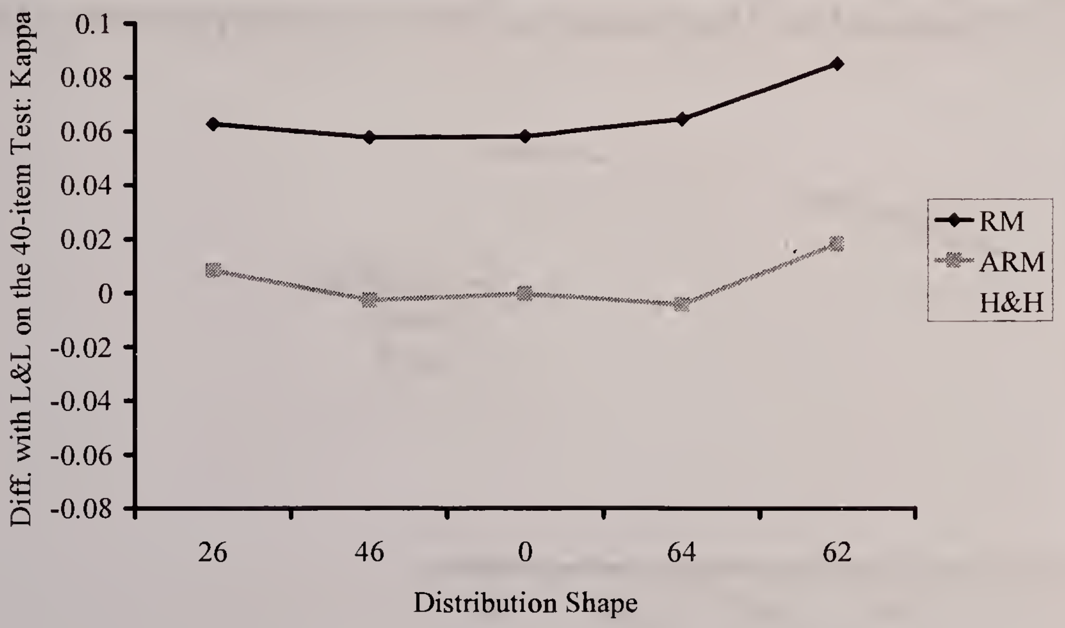


Figure 5.2.4.2.3 Difference with L&L for the 40-item Test: Kappa.

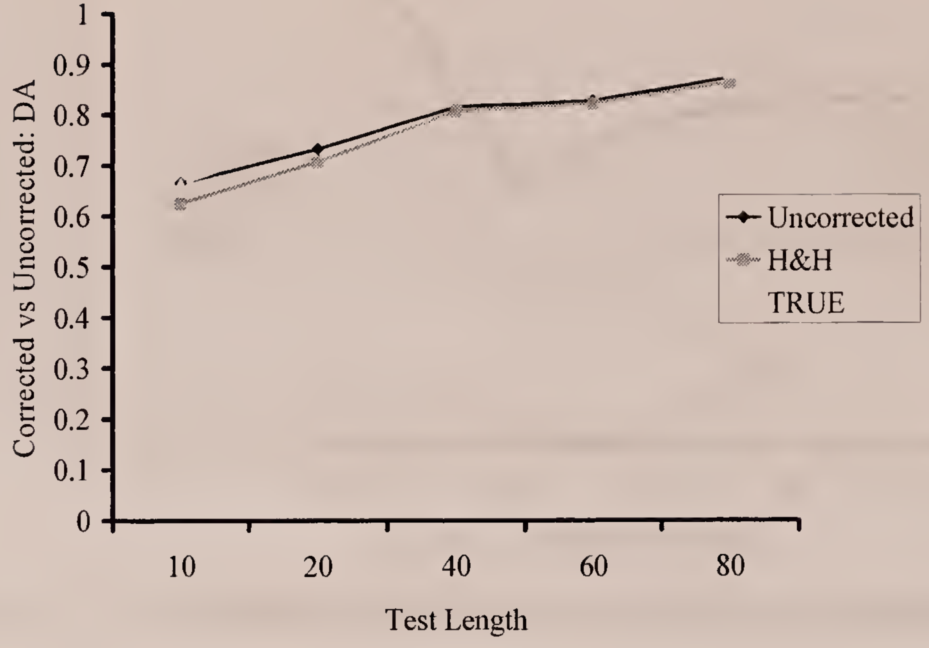


Figure 5.2.5.1.1 Corrected vs. Uncorrected on H&H: positively skewed.

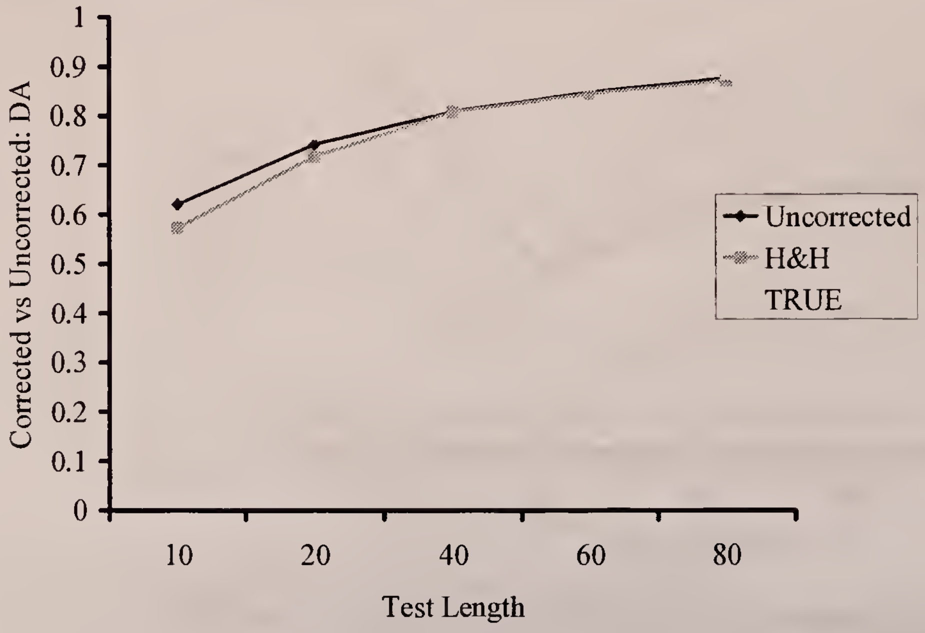


Figure 5.2.5.1.2 Corrected vs. Uncorrected on H&H: slightly positively skewed.

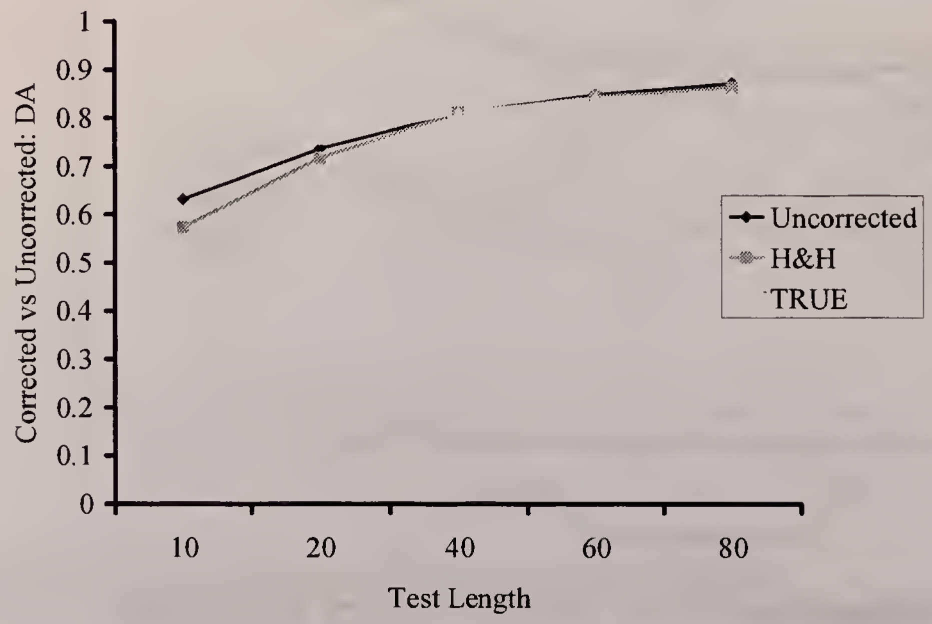


Figure 5.2.5.1.3 Corrected vs. Uncorrected on H&H: normal.

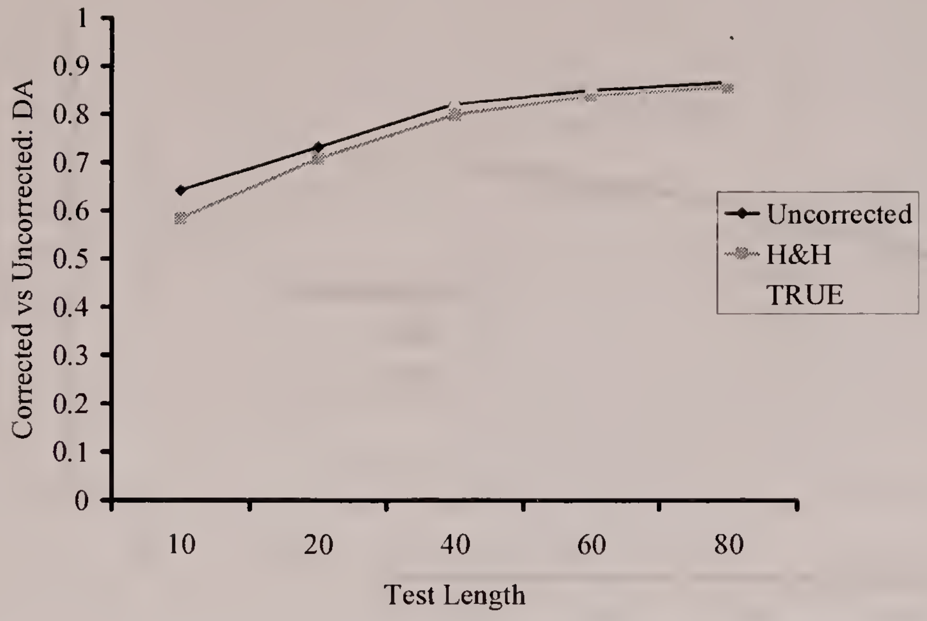


Figure 5.2.5.1.4 Corrected vs. Uncorrected on H&H: slightly negatively skewed.

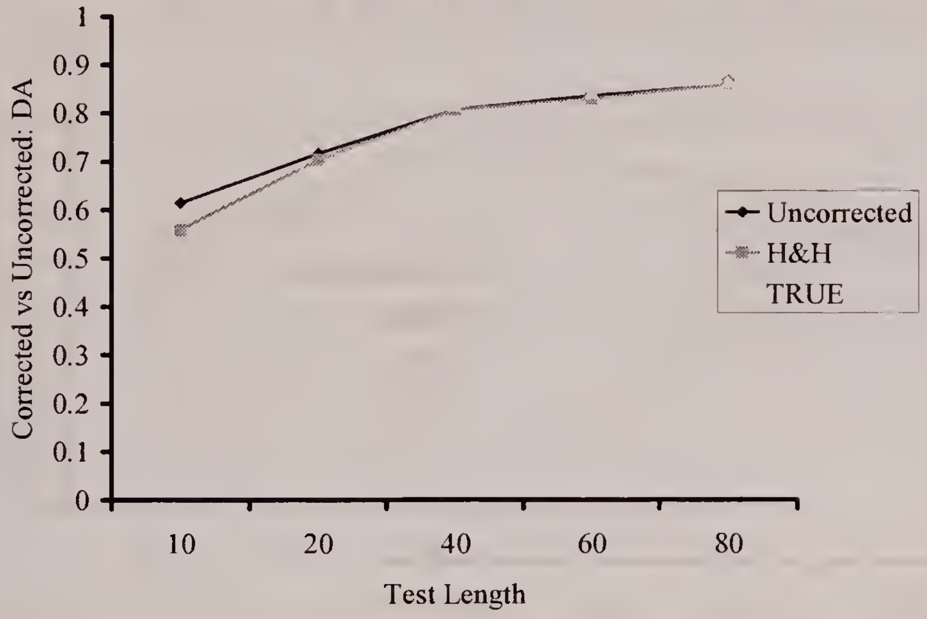


Figure 5.2.5.1.5 Corrected vs. Uncorrected on H&H: negatively skewed.

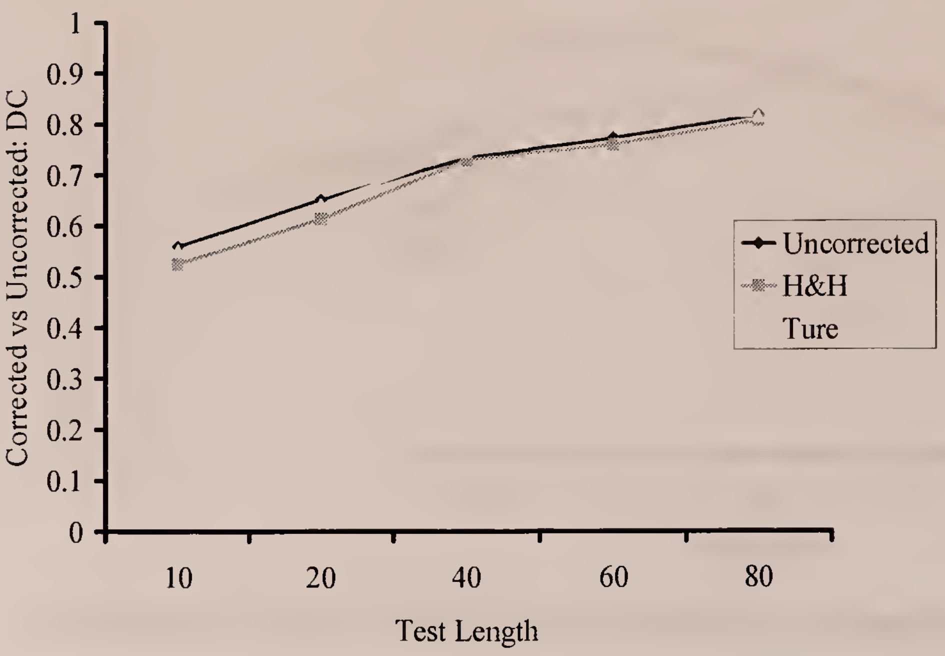


Figure 5.2.5.2.1 Corrected vs. Uncorrected on H&H: positively skewed.

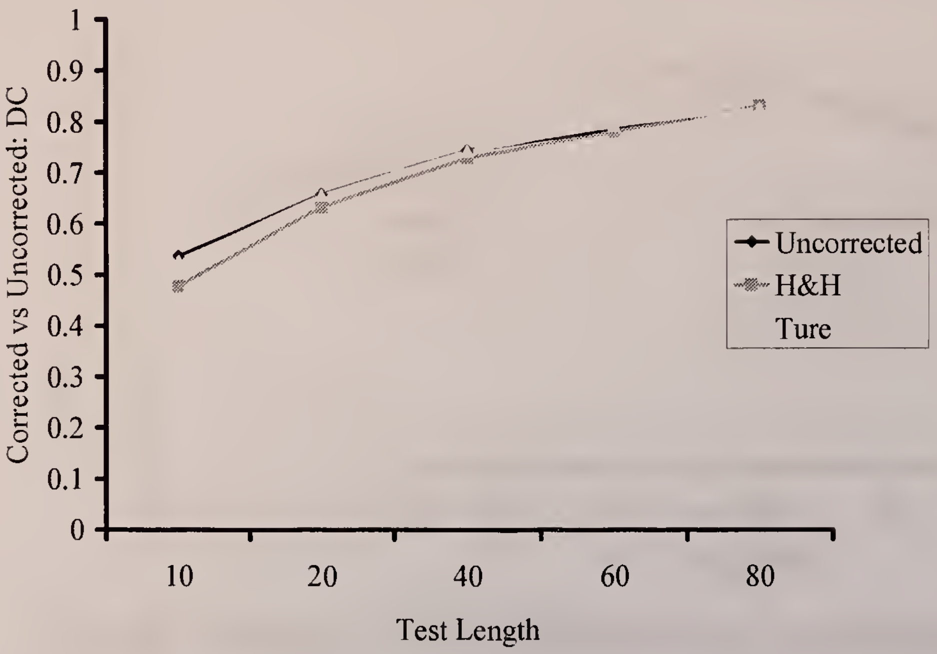


Figure 5.2.5.2.2 Corrected vs. Uncorrected on H&H: slightly positively skewed.

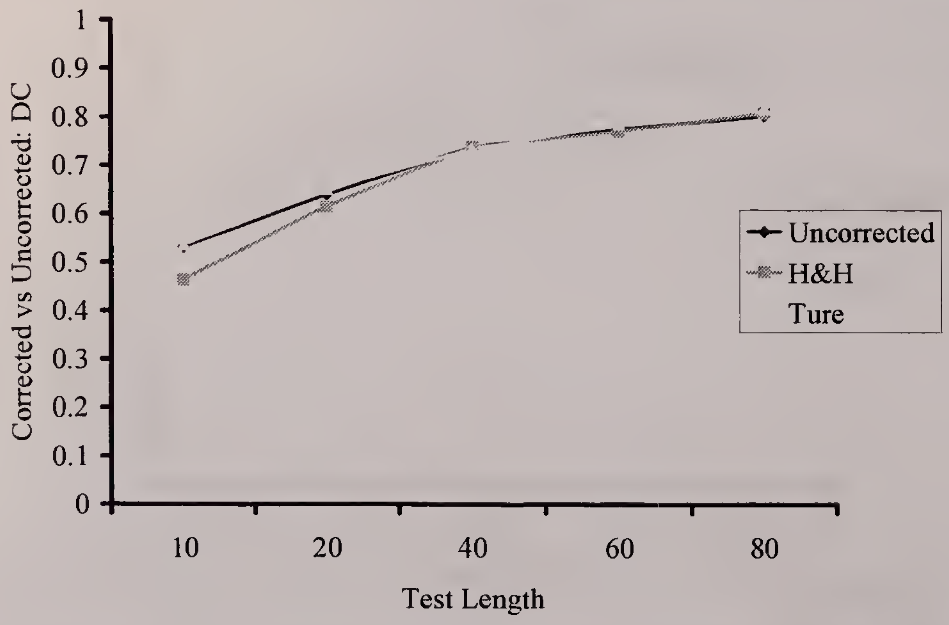


Figure 5.2.5.2.3 Corrected vs. Uncorrected on H&H: normal.

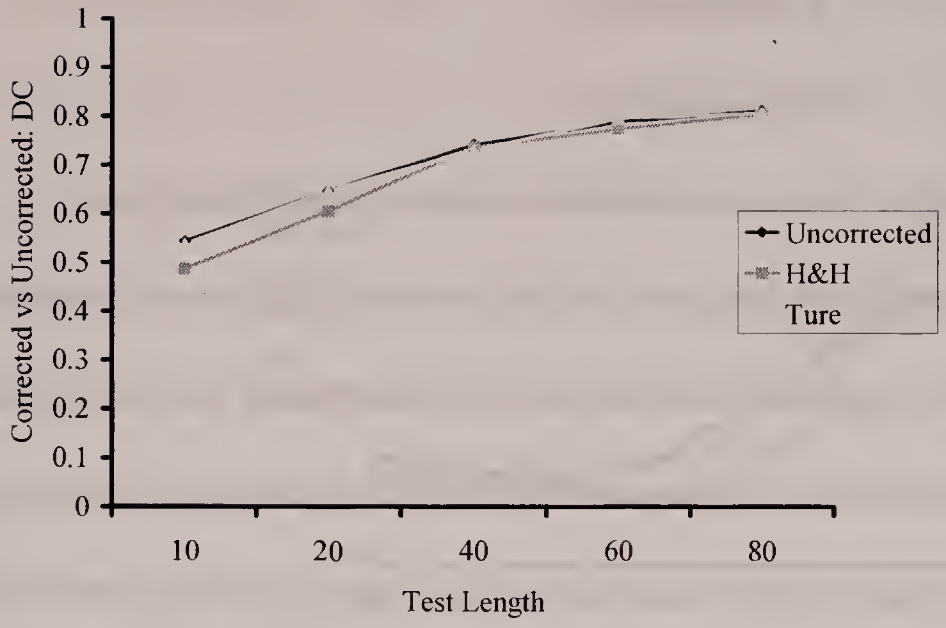


Figure 5.2.5.2.4 Corrected vs. Uncorrected on H&H: slightly negatively skewed.

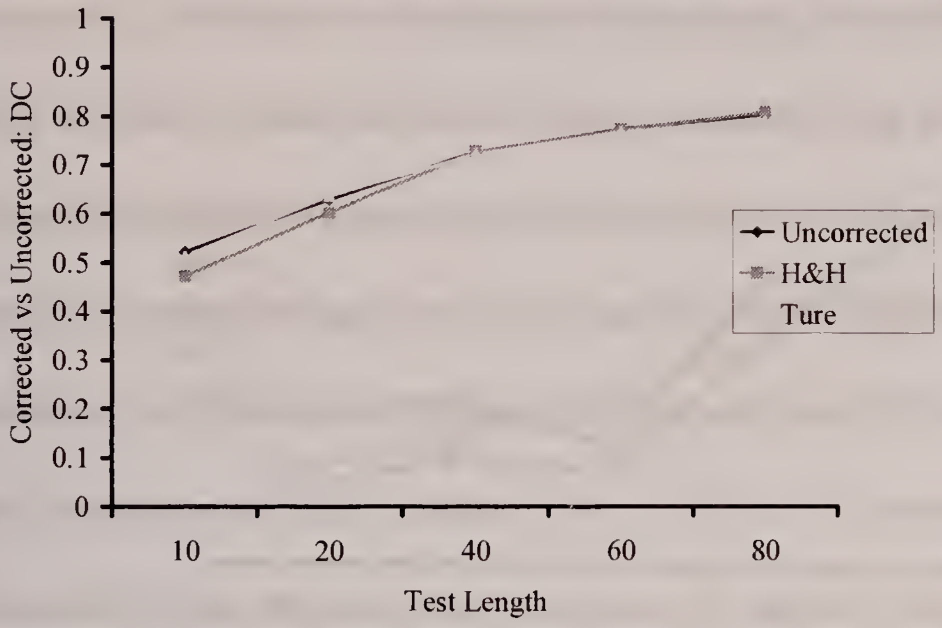


Figure 5.2.5.2.5 Corrected vs. Uncorrected on H&H: negatively skewed.

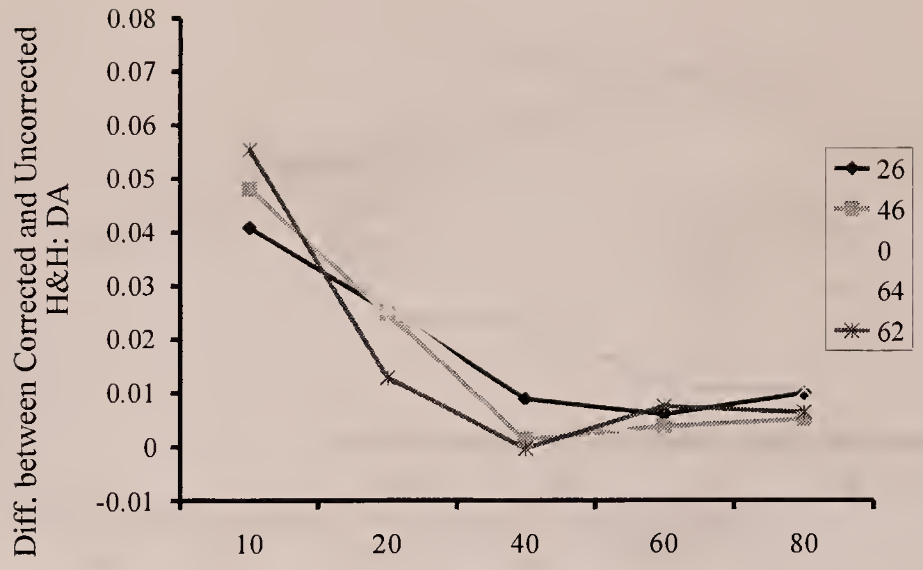


Figure 5.2.5.3.1 Difference between the corrected and uncorrected H&H: DA.

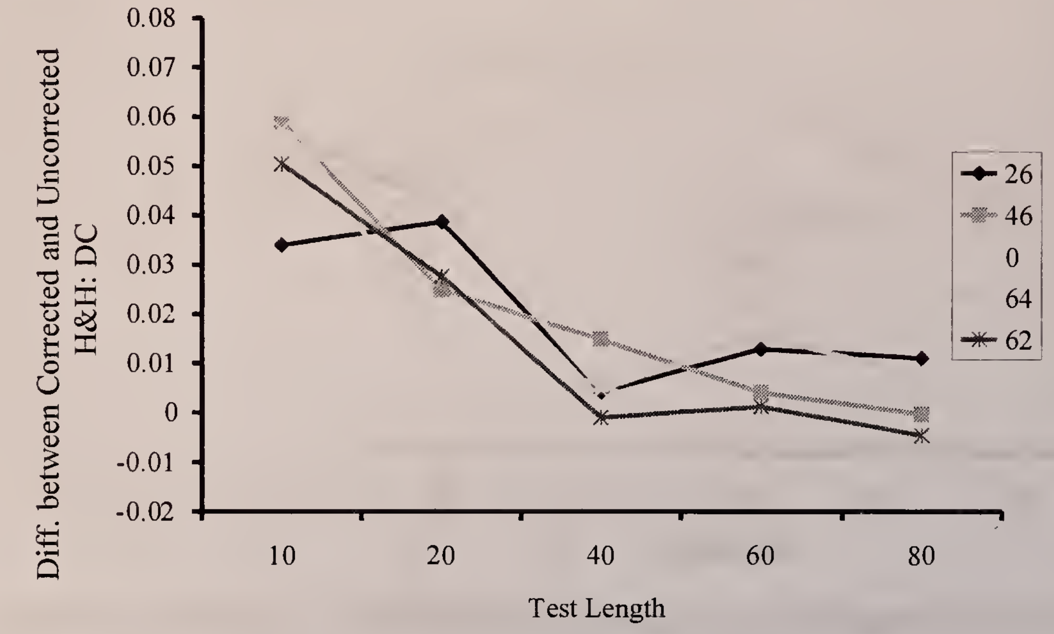


Figure 5.2.5.3.2 Difference between the corrected and uncorrected H&H: DC.

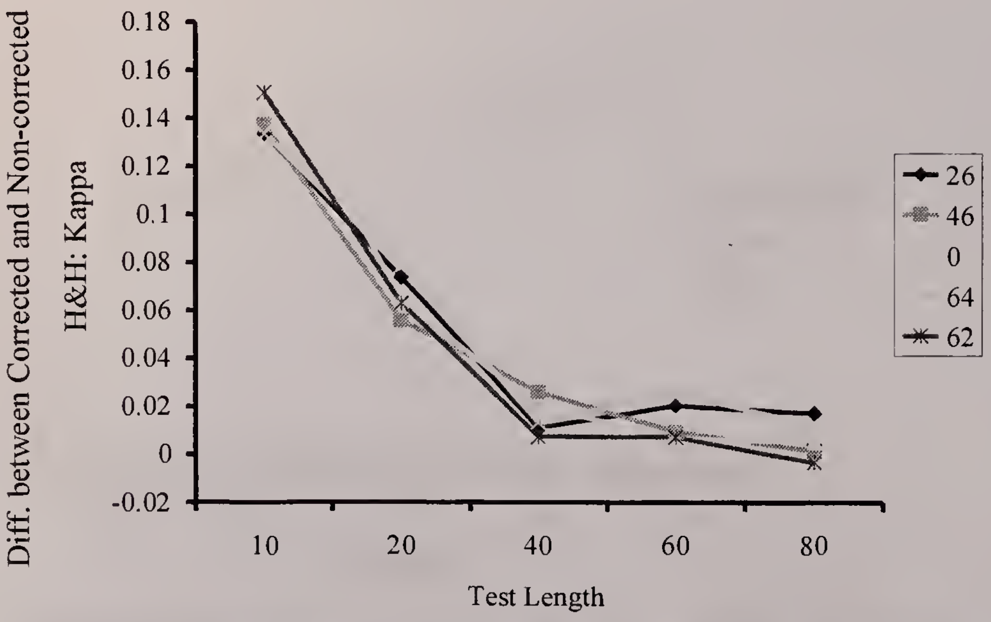


Figure 5.2.5.3.3 Difference between the corrected and uncorrected H&H: Kappa.

CHAPTER 6

DISCUSSION

In light of the high-stakes decisions involved in the increasing number of state mandatory tests, users of these kind of tests including test professionals, test publishers, researchers, and policy makers alike have to investigate thoroughly the psychometric properties of the tests and justify all the analysis procedures they apply on the tests. Given the critical role the decision accuracy and decision consistency are playing in high-stakes decision-making, a lot of attention has been turned to the approaches to estimating classification accuracy and consistency for proficiency scores. However, most of these methods are based on the classical testing theory, while the topic of estimating classification accuracy and consistency in the framework of IRT has been a relatively underdeveloped area of research. The comprehensive literature review conducted as a part of this project revealed only two IRT methods: Rudner (2001, 2004), and Hambleton and Han in Bouque et al. (2004). For some specific reasons as elaborated in the literature review section, the Rudner method was for DA only and H&H method was applied to the tests containing dichotomous data only. Given the importance of IRT in the field, this project aimed at introducing some IRT methods to evaluate DA and DC. The following discussion will summarize the findings from the study, develop the implications of these findings, and introduce several areas for further research.

The purpose of this project was to explore some possible methods for evaluating DA and DC in IRT. It first explicitly delineated 3 sets of IRT methods including their mathematics forms and the procedures for carrying them out in practice. Specifically, it extended the existing Rudner method to the estimation of DC and this represented the first set of IRT method in this project: RM. Based on the RM and some of its restrictions, it put forward the second set of IRT method: ARM. These two tasks were accomplished in Chapter 3 of this paper. Chapter 4 was devoted to the extension of Hambleton and Han method to test containing polytomous data which constituted the third set of IRT methods investigated in this project: Hambleton and Han. Then a series of simulation studies was carried out to evaluate the robustness of these methods under various conditions in Chapter 5. The design of the simulation studies resulted in 150 conditions that were evaluated using simulation studies with 5,000 examinees based on a statewide mandatory mathematics test for a certain grade. The efforts of evaluating these methods included comparing the results from the methods against the generated "true" DA, DC and Kappa values, comparing the methods to the commonly used CTT method L&L, and comparing the methods among themselves across 5 different test lengths and 5 different true score distribution shapes.

Chapter 3 and 4 detailed the development and implementation of the 3 IRT procedures involved, and illustrated how these methods could be applied to a set of real data. The results in these two chapters seemed to suggest that the three IRT methods produced reasonable and consistent values: the resultant DA values from RM, ARM and H&H on this set of data were all similar (0.8161, 0.7982 and 0.7986, respectively), as were the DC values (0.7408, 0.7211 and 0.7162, respectively), and Kappa values (0.6380, 0.6137 and 0.5893, respectively).

Chapter 5 resulted in 150 sets of DA, DC and Kappa values. Overall, these values all look reasonable in their own corresponding simulated conditions. As shown by Table 5.2.1, Table 5.2.2 and Table 5.2.3, all the first 125 sets of values strictly follow the pattern that DA values were greater than DC, and DC values were greater than Kappa values, as would be expected. Also, regardless the shape of true score distribution for the test forms, DA, DC and Kappa values all increased as the test length increased.

The 3 IRT methods and L&L were collectively evaluated by comparing their results to the "true" DA, DC indices generated on all simulated test forms. Overall, the absolute differences (AD values) between the results from these 4 methods and the "true" values on most of the simulated test forms were very small (within or around 3%). Specifically, with RM, no AD values were greater than 5% for either DA or DC, and only 2 AD values for Kappa were larger than 5% (on the positively skewed 10-item test and the negatively skewed, 80-item test, respectively). Regarding DA with ARM, all large AD values (larger than 5%) were displayed on the 10-item tests of all distribution shapes, with the AD values for the rest test forms being very small (within 3%). It had no greater than 5% values for DC, and only 2 such numbers for Kappa which happened on the positively skewed, 10-item test and the positively skewed, 60-item test, respectively. As to H&H, all greater than 5% AD values were happened on the short test forms. On the 10 short test forms (including all the 10- and 20-item tests), this method yielded more big values for DA (6 altogether) and Kappa (10 altogether) than for DC (2 altogether). With respect to L&L, again, all greater than 5% AD values were displayed on the short tests: 8 for DA, 2 for DC and 4 for Kappa.

The above results seem to suggest the following: (1) All 4 methods are sound methods, and are promising on most of the simulated test forms, since they produced small AD's most of the time. (2) Most of their big AD values occurred on the 10-item test forms and always with skewed true score distribution shapes. We might be able to conclude that none of the methods was proved to be robust on 10-item test with skewed true score distribution. At the same time, we should bear in mind that the total raw scores for these 10-item tests were only 16 points, and the reliability coefficients for all of them were smaller than .70. (3) It seemed the 3 IRT methods were more accurate than L&L, since they yielded slightly smaller and fewer AD values overall. But we should be cautious in drawing this conclusion, since it might be due to the fact that the simulation study was designed and carried out in the framework, and this might have been advantageous to the 3 IRT methods. (4) Among the 3 IRT methods, RM and ARM produced slightly closer results to the "true" values than the H&H, especially on the tests of shorter length. (5) RM and ARM tended to produce slightly higher values, with RM ones being the highest, while L&L and H&L producing some lower than "true" values.

As an important way to check on the consistency of the results derived from the 3 IRT methods, outcomes associated with each of the IRT method were also evaluated and compared to those from the L&L. The comparison results showed the following: (1) In general, all 3 IRT methods produced results that were close to those from L&L, so when the overall results from these methods were plotted in Figures 5.2.2.1.1 to 5.2.2.1.3, Figures 5.2.2.2.1 to 5.2.2.2.5, and Figures 5.2.2.3.1 to 5.2.2.3.5 the 4 lines representing the 4 methods were pretty tightly bundled for the most part and followed the same pattern for DA, DC and Kappa. This was expected since they were all close to the "true" results in the previous comparisons. (2) The results from the four methods displayed the most variability on the test that has the negatively skewed and 10 items. Recall that this test had total raw score of 12 and a reliability coefficient of 0.67. (3) As the tests grew longer, the difference between the methods grew smaller, especially beginning from the 40-item test. This was pretty noticeable in the figures mentioned previously. Recall that all 40-item tests had a reliability of around 0.91 and a total raw score of 64. It seemed this represented a condition where the results from all methods began to converge. (4) RM consistently produced the highest values, L&L consistently the lowest, with ARM and H&H in between. The author conjectured that it might due to the fact the RM operates on the theta metric, while the other 3 operate on the test score metric. More research is needed to explore the reason and nature of this method.

To better reveal the effect test length has on estimating DA, DC and Kappa using all the methods concerned, only the normal true score distribution shape (Shape 0) was considered across the 5 levels of the test length variable: 10, 20, 40, 60 and 80 items. The main findings from this analysis was as follows: (1) As expected, the resulted DA, DC and Kappa values grew in increasing order of test length, and the growth in DA, DC and Kappa indices from all methods was more drastic from the 10-item test to the 20-item test and from the 20-item test to the 40-item test than the growth from the 40-item to the 60item or from the 60-item to the 80-item test. (2) In this situation, most AD values between the results from all 4 methods were very small, except for the fact that ARM, H&H and L&L had a slightly greater than 5% value on DA on the 10-item test. (3) Compared with L&L, all IRT methods RM, ARM, and H&H produced closer DA, DC and Kappa results across the 5 different test lengths than when other shapes were also considered. (4) All the differences in (2) and (3) decreased as test length increased, for example, ARM and H&H had very similar differences with L&L once the test was at least 40-item long. These findings indicate that when examinees' true score distribution was close to normal, all methods were robust on tests that are at least 40-items long, that they even showed some robustness as well as instability on tests as short as having 10 items which had a total score of 12, that the results from the 4 methods displayed some variability on short tests that had 10 or 20 items, and that IRT methods produced results that were closest to the "true" values. However, again, we have to bear in mind that the simulated "true" values were generated in the framework of IRT.

The effect of the examinee true score distribution shapes on the estimation of DA and DC values for these methods was also checked by taking into consideration the only 40-item test, since it was closest to the real life test on which the simulation study was based. The highlights of the analysis results were as follows: (1) When compared to the "true" values, the resulted AD values seem to suggest that all 4 methods function very well across all true score distribution shapes, since no method had an AD value that was greater than 5% for DA, DC or Kappa. (2) All the AD values for the normal distribution were very small. They are generally smaller for RM and ARM than L&L and sometimes than H&H results, as would be expected, given that the two methods have an assumption that examinees' true score distribution is normally distributed. (3) Among the IRT methods, roughly, H&H seemed to yield the closest results to those from L&L.

The above findings from all the analyses for this research are encouraging. They provided evidence that would support the use of the 3 IRT methods introduced in this project in estimating DA and DC indices in most of the simulated situations, and in most of the cases the 3 IRT methods produced results that were close to the "true" DA and DC values, and consistent results to (sometimes even better results than) those from the commonly used L&L method. So far as the factor test length is considered, if the examinee score distribution was close normal, the 3 IRT methods produced very close results to "true" values on 40-item or longer tests, and sometimes on short tests such as 10-item ones and 20-items ones, but the results were not very stable across methods. With respect to the examinees' ability score distribution, the 3 IRT methods unanimously produced results that were very close to the "true" values on all the simulated test forms having different degrees of skewedenss. It seems the IRT methods showed more robustness on the distribution shapes than on the test length. When these two factors were mixed and taken into consideration at the same time, it required the 40-item long test (having a total raw score of 52) for all of the 3 IRT methods to produce stable results.

Another important advantage for these 3 IRT procedures is that they have proven to be mathematically simple and computationally efficient, compared with other methods. For example, for most of the previously existing CTT methods to produce one-administration DA and DC values, their benefit has been evident and convincing, but they are often based on complicated mathematical models, and there is always no readily available software for users to implement the procedures. The methods that rely on techniques such as bootstrapping can be theoretically sound and very accurate, but they are often costly and labor intensive computationally.

Of course, these methods are not without their own problems. RM and ARM, for example, rely on the assumption that students true scores are normally distributed (often accomplished by chosen a normal density with zero mean and unit variance). Also, the results from H&H on the same dataset vary slightly each time because it relies on simulation. On one hand, we can argue that in the case of RM and ARM, although it is known that a model may be inappropriate if one or more of its assumptions are violated, unfortunately, in practice, no model will ever have any of its assumptions satisfied strictly. Therefore, the important question that should be asked is not whether assumptions are satisfied, but at what point do the violations impact the proposed interpretations and uses severely, i.e., render the results invalid (Wells, 2004). In the case of H&H, we can overcome that slight problem by replicating the calculations for a number of times and take the means for the indices we want. On the other hand, the results of this project showed the robustness of the violation of normality under most conditions. The encouraging results from these IRT methods are motivation enough to further refine the methods. One important direction for further research is to investigate some feasible and efficient ways to estimate true score distributions.

The focus of the study is to introduce some IRT methods for evaluating DA and DC and to evaluate their performance in a series of reasonable conditions, but the results of this research also provide some information about relationship of the DA and DC indices of tests and some factors that might have an impact on them. For example, on the trend and variability of the relationship between DA and estimation accuracy, part of our results were consistent with the Ercikan and Julian (2002) findings that given a test with 4 proficiency levels, the test has to have a reliability estimate of around 0.90

to derive a DA value of around 0.80. Though this is not the intention of the current research, it will be very important to study how the methods for estimating DA and DC (including both IRT and CTT methods) can be influenced by some other factors such as measurement accuracy, proficiency levels and so on. The results from such a study can be used as a guideline for making decisions about test length and the number of proficiency levels in the design stage of an assessment, as well as in interpreting test performance results in terms of proficiency levels (Ercikan & Julian, 2002).

Given the special characteristics of two popular forms of testing: computer-adaptive testing and multi-stage testing, neither the Rudner nor the Adapted Rudner can be applied directly to these cases. However, the H&H method has very important implications in evaluating DC and DA with these kind of tests, and this possibility can be further explored.

Finally, it should be noted that in the calibration and simulation of all the test forms in this study, 3PL model for dichotomous data and GRM model for polytomous data were used. So the results might be specific to these models in a way and more study is needed to make the findings general to other models.

In summary, DA and DC have been playing a critical role in the educational measurement field as well as all licensure exams. The review of the literature clearly points to the importance of this topic and all the important and influential results from the previous attempts at estimating DA and DC. At the same time, it has also revealed that more work in this area is warranted in estimating DA and DC in IRT. The major purpose of the research is to explore some feasible IRT methods, and the findings from the simulation studies and the real data study suggest that the three IRT methods introduced in this research were somewhat successful in doing so, given that they produced close results to “true” values and to the L&L methods in most situations, and that they are at least not more complicated than the CTT methods mathematically or computationally. The information from this study also provides some useful information on the effects some factors may have on DA and DC estimation methods. This project represents the beginning of a continuous effort to establish feasible and refined IRT methods to estimate DA and DC and, and to accurately describe the relationship between DA and DC with the factors that have an impact on them so that this information can not only help the test users to derive accurate DA and consistent DC values, but also serve as guidelines for them to refer to before they build a test.

BIBLIOGRAPHY

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. A. (1980). A consumers' guide to criterion-referenced test reliability. *Journal of Educational Measurement, 17*(4).

Berk, R. A. (1984). Selecting the index of reliability. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 231-266). Baltimore: Johns Hopkins University Press.

Brennan, J., & Wan, L. (2004). *A bootstrap procedure for establishing decision consistency for single administration complex test.* Paper presented at the Annual Meeting of the National Council of Measurement in Education. San Diego, CA.

Bourque, M. L., Goodman, G., 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.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Crocker, L., Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Harcourt Brace Jovanovich College Publishers.

Embertson, S. E. & Reise, S. E. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

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.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist*, *18*, 519-521.

Hambleton, R. K., & Han, N. (in press). 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.* Washington: Degnon Associates.

Hambleton R. K, & Li, S. (in press). Criterion-referenced testing: Purposes, technical issues and advances. In B. Everitt & D. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science.* West Sussex, UK: John Wiley & Sons.

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.

Hambleton, R.K, & Traub, R. (1973). Analysis of empirical data using two logistic latent trait models. *Br. J. math. Statist. Psychol*, *26*, 195-211.

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 decision in domain-reference testing. *Journal of Educational Measurement*, *13*, 253-264.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika*, *27*.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating methods and practices*. New York: Springer-Verlag.

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

Livingston, S. A., & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement, 16,* 247-260.

Lord, F. N. (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.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equating." *Applied Psychological measurement*, *8*, 452-461.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177- 196.

Muraki, E., & Bock, R. D. (1986). PARSCALE: IRT Item Analysis and Test Scoring for Rating-Scale Data. [Computer program]. Chicago, IL: Scientific Software International, Inc.

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

Peng, C.-Y. 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.

Rudner, L.M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, *7*(14).

Rudner, L.M. (2004). *Expected classification accuracy*. Paper presented at the annual meeting of the National council on Measurement in Education. San Diego, CA.

Rulon, P.J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, *9*, 99-103.

Spray, J. A. & Welch, C. J. (1990). Estimation of classification consistency when the probability of a correct response varies. *Journal of Educational Measurement*. *27*(1).

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

Subkoviak, M. J. (1978). Empirical investigation of procedures for estimating reliability for mastery tests. *Journal of Educational Measuremen*t, *15*(*2*).

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

Swaminathan, H., Hambleton, R. K., & Algina, J. (1975). A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement*, *12*, 87-98.

Thissen, D., Pommerich, M. Billeaud, K., & Williams, V.S.L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*, 39-49.

Traub, R. E. (1994). *Reliability for the Social Sciences*. Thousand Oaks, California: Cage Publications.

Wells, C. S. (2004). Investigation of a nonparametric procedure for assessing goodness-of-fit in item response theory. Unpublished doctoral dissertation, University of Wisconsin at Madison.

Wilcox, R. R. (1981). A review of the beta-binomial model and its extensions. *Journal of Educational Statistics, 6,* 3-32.