THE NEW GOODNESS-OF-FIT INDEX FOR THE MULTIDIMENSIONAL ITEM RESPONSE MODEL

By

Shu-chuan Kao

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology, and Special Education

2007

ABSTRACT

THE NEW GOODNESS-OF-FIT INDEX FOR THE MULTIDIMENSIONAL ITEM RESPONSE MODEL

By

Shu-chuan Kao

The current research is concerned with the goodness-of-fit of the multidimensional item response theory (MIRT) model to binary test data. Based on the R^2 analog proposed by Estrella (1998) for the dichotomous dependent variable model, the new goodness-of-fit index, the *RLR* index (Ratio of Log Residuals), was proposed to reflect the ratio of error reduction when adding dimensions to the MIRT model.

The *RLR* index demonstrated nice statistical properties in term of the results from two simulation studies. Compared to the G^2 test and G^2 difference test from TESTFACT, the *RLR* index could identify true dimensionality with Type I error rates less than .05 and demonstrate high statistical power to reject wrong models for most cases. The findings also indicated that the *RLR* index was sensitive to different levels of item discrimination, the variation of item difficulties, inter-factor correlation, and item-factor structure. It was also found that a large sample size and a long test could generate more accurate dimensionality decisions. Regarding the analysis of real data, one statistical dimension was suggested to describe the Grade 4 Mathematics Test of the Michigan Educational Assessment Progress (MEAP) testing program. The unidimensional finding was supplemented with the discussions in term of the test item content, the representativeness of the content-related dimensions, the definition of dimensionality, and the assumptions of the compensatory MIRT model.

ANKNOWLEDGEMENTS

There are a number of people who have helped me to this point, and whom I would like to acknowledge. First, I wish to express my sincere gratitude to my advisor and dissertation chair, Professor Mark Reckase. He has held my hand throughout my pursuit of Ph.D., serving as both a good friend and an advisor. His guidance and support not only made this dissertation possible, but also enabled me to become a psychometrician. I would also like to thank my dissertation committee members: Professor Richard Houang, Professor Yeow Meng Thum, and Professor Alexander von Eye. Their comments, suggestions, and encouragement not only made for a stronger dissertation but also helped me get a better view of research. Special thanks should go to my friends: Jane Lin, who has been my C language tutor for the past 3 years, gave me unwavering support and helped me correct my malfunctioning pointers; Jules and Andy, who have been my good friends and mathematics tutors, patiently taught me the essence of mathematics; Kang-Hung, who has always been willing to share his expertise in econometrics with friends, helped me find a different approach to investigate dimensionality; Yun-Jia, who has been the source of encouragement for the past 3 years, helped me collect laptops to run TESTFACT and find resource to finish this study. Besides, I have to thank all those friends who have been around me and brought light to my life. Finally, my deepest appreciation goes to my family, who is my lifelong support and has given me freedom to fulfill my dreams.

My experience of doing this dissertation is quite rewarding. Without the pain, there won't be the harvest. Without conducting this research, I won't have a chance to know that these professors are so smart and warm, and my friends are so lovely and supportive.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Dimensionality plays an important role in test score interpretation and the validity of inferences made from tests, and is one of the critical issues in educational measurement. For many testing practitioners, it seems unreasonable to use the common data analysis procedures assuming that the data are unidimensional while the assessment tools, especially achievement tests, are designed to measure multiple content knowledge and skills. When tests are planned to measure different cognitive abilities or content knowledge, and examinees are required to demonstrate more than one ability to answer items correctly, the properties of the resulting test response data are difficult to describe. For instance, a mathematics test may contain "story-type" questions. From the psychological point of view, examinees will have to use mathematical skills and reading abilities to correctly answer such questions. From the statistical point of view, psychometricians may need more than one statistical variable to represent each person in order to sufficiently model the interaction between test items and examinees. Describing the statistical characteristics of potentially multidimensional data by the traditional procedures assuming unidimensionality may not only cause measurement problems but also lead to inaccurate score interpretation.

1.1 Different Perspectives to Investigate Data Dimensionality

With the intention to investigate the likely multidimensional nature embedded in the item response data, psychometricians have developed different perspectives to interpret

dimensionality. Based on Embretson's (1985) definition, dimensionality indicates the number of hypothesized psychological constructs required for successful performance on a test. This definition of dimensionality can be referred to as "psychological dimensionality." In psychological measurement, the number of dimensions in the model is often based on cognitive theories and each dimension represents a specific latent trait being modeled. In educational testing, the psychological constructs are often attributed to content domains of interest, reflecting the purpose of the test. However, in the real testing situation, the sources of multidimensionality are still unclear. Besides the desired psychological traits or content knowledge, other undesirable factors that may be the cause of multidimensionality include: different item format (Tate, 2002); test speededness (Bock, Gibbons, & Muraki, 1988; Douglas, Kim, Habing, & Gao, 1998); item dependency from testlet items (Ferrara, Huynh, & Michaels, 1999; Thissen, Steinberg, & Mooney, 1989); and inappropriate design of test administration conditions (Tate, 2002). Determining the number of psychological dimensions to model test data, or deciding how well the model fits data, requires validity studies to supplement the statistical index. This implies that even if the test is known for requiring examinees to demonstrate two different cognitive abilities to answer the items correctly, validation studies are needed to verify that the two psychological dimensions in the model match the hypothesized constructs.

Another definition of dimensionality is based on the statistical properties of the test data. According Lord and Novick's (1968) definition, dimensionality is the total number of abilities required to satisfy the assumption of local independence. This assumption indicates that an examinee's responses to the items in a test are statistically

independent if their ability level is taken into account. The probability of any particular item response pattern for an examinee is the product of individual item probabilities. When the assumption of local independence is satisfied, the complete latent space is defined and, at the same time, the number of dimensions needed to summarize the data is specified. In terms of these explanations, this kind of definition of dimensionality can be referred to as "statistical dimensionality."

Unlike the psychological dimension, determination of the number of statistical dimensions depends on the mathematical properties in the data under the assumption of local independence and monotonicity[1]. Harrison (1986) and Tate (2002) concluded that every set of test responses is multidimensional to some degree. To decide the data dimensionality, many researchers (Berger & Knol, 1990; Junker & Stout, 1994) suggested that the latent traits that underlie test data can be classified as major (i.e., dominant) and minor factors. Humphreys (1985) argued that the construction of tests that are valid for intended purposes requires tests that are sensitive to differences on a dominant trait and numerous minor factors. In order to measure the dominant factor of interest (e.g., computation ability), the inclusion of numerous minor factors is inevitable. Wainer and Thissen (1996) suggested that item responses will always reflect either random or fixed multidimensionality. The random multidimensionality is caused by the presence of minor dimensions or nuisance dimensions other than those planned to determine the responses. The fixed multidimensionality corresponds to the number of dimensions the test is designed to measure. Concerning the unidimensionality assumption of IRT, Ackerman (1994) pointed out that the unidimensionality should never

1 Suppes and Zanotti (1981) proved that all the data can be modeled unidimensionally when the restriction of monotonicity is relaxed. In this case, the dimensionality is no longer an issue in data modeling. However, the explanation of the relationship between ability and item response will be obscure.

be assumed but should be verified. It would be considered problematic to analyze multidimensional data with the statistical procedures assuming that the data are unidimensional.

To clarify the connections and distinctions between psychological and statistical dimensions, researchers (Reckase, 1990; Reckase, Ackerman, & Carlson, 1988) defined dimensionality as the minimum number of mathematical variables needed to summarize a matrix of item response data. In other words, to fully describe all the differences related to the test for the examinees in the population, the minimum number of statistical abilities required in the model would be considered as test dimensionality. Reckase (1990) indicated that for a test to be modeled unidimensionally, tests do not have to measure narrowly defined, pure psychological traits for statistical procedures that assume unidimensionality. Test items that measure the same combination of traits will likely generate unidimensional data when examinees interact with them. Therefore, it is possible to have statistically unidimensional item response data even though the psychological dimensions needed to correctly answer the questions are greater than one.

1.2 Dimensionality and Multidimensional Item Response Theory

Determining the number of dimensions needed to explain the item response data is often of substantive or methodological interest not only for educational measurement, but also for psychological studies. Spearman (1904) first argued that the performance on sets of tests could be explained by individuals' levels on general and specific traits. Since then, determining the number of dimensions needed to summarize a set of data has been an important research question. The study of test dimensionality is the essential

issue for the investigation of test construction, test validity, reliability, fairness, and the interpretation and use of test scores (Choi, 1997; Tate, 2002). For the past decades, a number of studies have been conducted to explain test data relaxing the restriction of unidimensionality assumption, and the methodology of the Multidimensional Item Response Theory (MIRT) has been more widely accepted. MIRT offers a new methodology to analyze test data in such an elaborate way that item characteristics are independent of the sample, and the examinees' ability estimates are not test-dependent. However, the appropriate use of any MIRT model depends upon the good fit between model and data. All the MIRT-related testing techniques, such as multidimensional parallelism, multidimensional equating, multidimensional-based computerized adaptive testing, can be performed only when the data dimensionality is specified. Thus, it can be concluded that the applicability of MIRT rests on the availability of an appropriate model-data-fit index.

Beyond generating different mathematical MIRT models, researchers also proposed various model-data-fit indices to help determining the appropriate number of dimensions used in the MIRT models. However, no procedure for MIRT model selection has been universally accepted so far. Even though the MIRT calibration computer programs, such as TESFACT (Wilson, Wood, Gibbons, Schilling, Muraki, & Bock, 2003) and NOHARM (Fraser, 1988), are available, the problem of deciding the number of dimensions needed to model the data is still very much a topic of investigation. The current goodness-of-fit indices (e.g., the G^2 test provided by TESTFACT and the indices based on residual analysis) do not demonstrate good statistical properties in dimensionality detection (Berger & Knol, 1990; De Champlain & Gessaroli, 1991;

Hambleton & Rovinelli, 1986; Mislevy, 1986). In order to correctly analyze test data with MIRT, the development of a valid model-data-fit statistic is not only desirable, but necessary.

1.3 Purpose of the Study

The main purpose of this study is to propose and assess the use of the new goodness-of-fit index for MIRT model selection. Specifically, the degree to which the minor factors should be considered significant was evaluated in terms of the proposed index. Based on the results of simulation studies, the research demonstrated the accuracy and stability of the proposed goodness-of-fit index in detecting true dimensionality of test data under various testing conditions. The statistical characteristics of the proposed index were compared with those of the traditional χ^2 tests. Besides demonstrating the statistical properties for the simulated data, real test data were used to show the applicability of the proposed index in a real testing situation.

The significance of the study is to offer a more reliable and testable goodness-of-fit index with which to determine the number of dimensions for the MIRT model to properly calibrate test data. The procedure proposed in this study offers the theoretical base and empirical evidence to decide the goodness-of-fit for MIRT models. The results of this work have potential use for both theoretical researchers and those who work in applied measurement. With this information, MIRT users would have better reference to decide the minimum number of dimensions needed to model test data and make more valid use of test theories.

CHAPTER 2

LITERATURE REVIEW

To begin this chapter, the MIRT model used in this study is elucidated in detail. The chapter then provides a review of model-fit studies concerning MIRT. Next, a new goodness-of-fit index is proposed along with the theoretical background. Finally, evidence is presented to demonstrate the feasibility of applying the index to describe the model-data-fit for MIRT model.

2.1 Multidimensional Item Response Theory

Psychometricians have developed a number of MIRT models (see Reckase & McKinley, 1982; van der Linden & Hambleton, 1997) assuming a specific form of the item-examinee interaction on the basis of more than one ability dimension and attempt to decide the number of dimensions and which item measure which dimensions. Classified by their mathematical forms, these models can be distinguished as compensatory or partially compensatory, that is, whether or not high ability on one trait can compensate for low abilities on other traits. For the compensatory models (e.g., McDonald, 1967; Reckase, 1985; Reckase & McKinley, 1991), the performance on the item is determined by a linear combination of the multiple abilities so that high ability on one dimension can compensate for low abilities on other dimensions. By having high abilities on some dimensions, a probability of 1 for correct response can be obtained even with very low abilities on other dimensions (Reckase, 1997b). Concerning the partially

compensatory models (Sympson, 1978; Whitely, 1980)[2], the probability of correct response decreases with an increase in the number of dimensions (Reckase, 1997b). The multiplicative nature of the model allows an examinee to partially compensate for low abilities on one dimension by being high on other dimensions. Because most of the research on dimensionality has been done using compensatory models and the calibration computer programs are currently available only for that model, the logistic multidimensional compensatory two-parameter IRT model (Reckase, 1985; Reckase & McKinley, 1991) was employed in this study.

In this model, the probability of a correct response to item *i* can be expressed as

$$P(u_{ij}=1|\vec{a}_i,d_i,\vec{\theta}_j)=\frac{\exp(\vec{a}_i'\vec{\theta}_j+d_i)}{1+\exp(\vec{a}_i'\vec{\theta}_j+d_i)}, \quad (1)$$

where $P(u_{ij}=1|\vec{a}_i,d_i,\vec{\theta}_j)$ is the probability of a correct response of person *j* on item *i* in the *k*-dimensional ability space,

u_{ij} represents the item response for person *j* on item *i*,

$\vec{a}_i$ is a vector of parameters representing the discriminating power of item *i*,

d_i is a parameter related to the difficulty of item *i*,

$\vec{\theta}_j$ is the vector of abilities for examinee *j*, and,

e is the mathematical constant 2.7183.

Under this framework, each examinee is represented as a data point in this *k*-dimensional latent space. This equation defines a surface indicating that the

---

[2] For example, Sympson's (1987) model can be expressed as

$$P(X=1|\vec{\theta}_j,\vec{a}_i,\vec{b}_i)=\prod_{k=1}^{n}\left[1+\exp[a_{ik}(\theta_{jk}-b_{ik})\right]^{-1},$$

where *k* indicates the dimension; a_{ik} and b_{ik} are the discrimination and difficulty parameters, respectively. The root of the second derivative of this equation does not define a difficulty function but gives a single value for each dimension. That is, there is *b* parameter for each dimension.

probability of a correct response for a test item is a function of an examinee's location in the ability space specified by the θ-vector. The elements of the θ-vector are statistical constructs that may or may not correspond to any particular psychological traits or educational achievement domains (Reckase, 1997a). Besides, there is nothing in the model that requires the θ-coordinates to be uncorrelated. The θ-coordinates are for orthogonal axes, but the coordinates may be correlated. If the correlations among the θ-coordinates are constrained to be 0.0, then the observed correlations among the item scores will be solely accounted for by the discrimination parameters (Reckase, 1997a).

The interpretations of the model parameters are somewhat different from those in the UIRT model. The item discrimination parameter for the MIRT model, assuming orthogonal axes, is represented by Reckase and McKinley (1991) as the length of the discrimination vector. The length, $MDISC_i$, as shown in equation (2), indicates the maximum overall item discrimination of the item i for the best combination of abilities. The computation of $MDISC_i$ can be expressed as

$$MDISC_i = \sqrt{\sum_{k=1}^{K} a_{ik}^2}\ , \tag{2}$$

where k is the number of dimensions in the θ space, and a_{ik} are elements in the vector a_i given in equation (1). The discrimination of an item is a function of the slope at the steepest point and is best in a particular direction in the multidimensional space. The direction of the greatest discrimination in the multidimensional space is

$$\cos\alpha_{ik} = \frac{a_{ik}}{MDISC_i}, \tag{3}$$

where α_{ik} is the angle from the k-th dimension.

The item difficulty parameter, $MDIFF_i$, is defined as

$$MDIFF_i = \frac{-d_i}{MDISC_i}. \tag{4}$$

This value indicates the distance from the point of best discrimination to the origin. $MDIFF_i$ can be interpreted much like the b-parameter in UIRT. A negative $MDIFF_i$ value suggests an easier item, whereas a positive value indicates one more difficult.

Graphically, test items can be summarized by a vector plot so that the geometrical characteristics of *MDISC* and *MDIFF* can be clearly represented. A two-dimensional example, as shown in Figure 2.1.1, shows that the distance from the vector's base to the origin is *MDIFF*, and the length of the vector is *MDISC*. The extension of the vector goes through the origin, and the base of the vector is located on the line where examinees have a .50 probability to answer the item correctly. The vector plot allows plotting more than one item on one graph. Item vectors pointing in the same direction measure the same combination of θ_1 and θ_2. By examining the directions of the item vectors, the similarities among items and the dimensional structure can be identified.

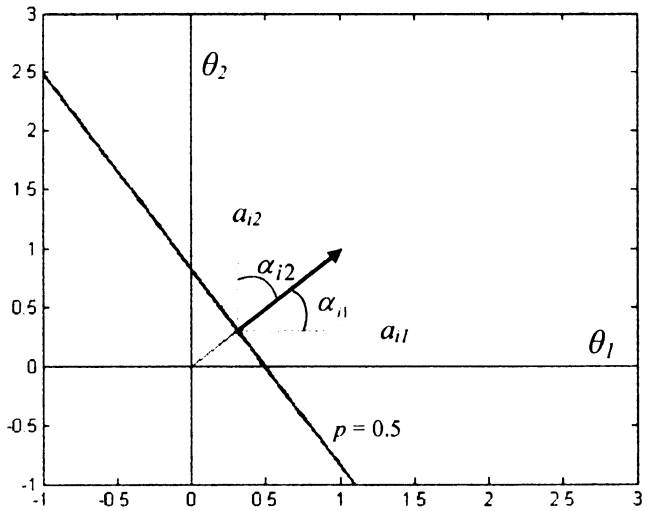


Figure 2.1.1. Item vector plot (a_1= 1, a_2= 0.6, d= - 0.5)

2.2 Review of Goodness-of-Fit Indices for Multidimensional Item Response Models

The dimensionality of test data is difficult to assess and is often based on personal judgment. Several studies (Berger & Knol, 1990; De Ayala & Hertzog, 1991; De Champlain & Gessaroli, 1996; Douglas, Kim, Roussos, Stout, & Zhang, 1995; Hambleton & Rovinelli, 1986; Hattie, 1984, 1985; Nandakumar, 1994; Roznowski, Tucker, & Humphreys, 1991; Stone & Yeh, 2006; Tate, 2003) were conducted to compare the relative effectiveness of the statistical procedures for detecting dimensionality of test data. These methods, available for assessing dimensionality, can be divided into two types: parametric and nonparametric procedures. The parametric procedure includes methods based on the mathematical equivalence between factor analysis models and the MIRT models (Knol & Berger, 1991; McDonald, 1967, 1985, 1989a). These studies suggested that the problem of assessing dimensionality in MIRT models for dichotomous data can be approached from a factor analytical point of view. An interpretation of multidimensional data structure is derived from the estimated factor loadings of the model. Conversely, the nonparametric procedure involves a collection of methods that avoid the problem of fitting an assumed parametric model[3]. The item covariance-based methods only assume that the item response function is monotonic and assessing dimensionality involves evaluating the conditional item associations. However, to perform goodness-of-fit studies, McDonald and Mok (1995) emphasized that the latent

[3] The item covariance-based methods include: Stout's essential unidimensionality procedure (Nandakumar & Stout, 1993; W. F. Stout, 1987) implemented in DIMTEST (W. Stout, Habing, Kim, Roussos, & Zhang, 1993); assessing multidimensional approximate simple structure DETECT (Kim, 1994; Zhang & Stout, 1995, 1996); hierarchical cluster analysis HCA/CCPROX (Roussos, 1992, 1993; Roussos, Stout, & Marden, 1998) based on proximity measure; Holland-Rosenbaum's test of unidimensionality, conditional independence, and monotonicity (Holland & Rosenbaum, 1986; Rosenbaum, 1984); Bejar's dimensionality assessment procedure (Bejar, 1980, 1988); and Tucker and Humphrey's methods on the principle of local independence and second factor loadings (Roznowski et al., 1991).

trait dimensionality should be assessed on the basis of the misfit of a latent trait model, not by indices that are not based on the model to be fit. Since this study only focuses on the compensatory logistic MIRT model, only the fit indices based on the parametric procedures, which can be classified into four types, will be included in the following sections. Even though different methods were proposed in the past, the focus of the problem was the same: to decide whether the minor factors are large enough to represent significant dimensions, or whether they are merely nuisance in the data.

2.2.1 Exploratory Linear Factor Analysis

Principal Component Analysis (PCA) and common Linear Factor Analysis (LFA) have been popular methods for exploring the dimensionality of dichotomous test data. In the studies of PCA or LFA, determining the number of components is often based on the amount of explained variance from phi or tetrachoric correlation matrices. Among the procedures are the well-known eigenvalue greater than 1.0 rule (Kaiser, 1960) and the scree plot test (Cattell, 1966).

The phi correlation coefficients generally produce a positive definite correlation matrix and tend to avoid the problem of Heywood cases (Berger & Knol, 1990). However, the LFA of phi correlation matrix was found to overestimate the number of underlying dimensions in any data (Hambleton & Rovinelli, 1986). The identification of spurious difficulty factors is related to the characteristics of the items rather than to true underlying relationships (Guilford, 1941). That is, the choice of cut points affects the values of the expected phi correlation coefficients. Factor analysis of phi correlation matrix of binary variables produced by the same underlying correlation structure but dichotomized at different cut points can conform to factor models with different structure

and different numbers of factors (Mislevy, 1986).

LFA of tetrachoric correlation matrix theoretically can avoid the problem of "difficulty" factors for dichotomous free-response items. Tetrachoric correlation coefficients can produce better estimates of the correlation than phi correlation coefficients, but the assumptions, such as the distribution of the latent variables being bivariate normal, and the latent variables being measured at the interval level should be obtained (De Ayala & Hertzog, 1991). However, when ability distributions are not normal and the item response function is not normal ogive, the use of tetrachoric correlations is inappropriate (Lord, 1980). Furthermore, tetrachoric correlation coefficients will become unstable when extreme values are reached. Tetrachoric correlation matrix will often not be positive definite and is more likely to produce Heywood cases (Berger & Knol, 1990).

Although the criticism of the use of tetrachoric correlation in LFA was clear, some researchers still found it useful when used appropriately. Knol and Berger (1991) considered various common factor analysis methods and concluded that, for large-scale applications, an unweighted common factor analysis of tetrachoric correlations performed as well as other techniques (e.g., full-information factor analysis). Drasgow and Lissak (1983) suggested that interpretation of data dimensionality could be enhanced by comparing the scree plot created from real data to that created from a factor analysis of randomly generated test data containing the same number of items. Ackerman (1994) concluded that these methods may sometimes be inconclusive and lead to spurious counting of dimensions, but the size of the eigenvalues in conjunction with a substantive review of the items can lead to the conclude of how many essential traits are being

measured.

2.2.2 Confirmatory Linear Factor Analysis

McDonald (1981) suggested that the factor analytic models of item response data can be tested with CFA, a technique often considered to be a special case of Structural Equation Modeling (SEM). McDonald and Mok (1995) asserted that the indices developed for SEM under the assumption of continuous variables could be applied to the assessment of dimensionality for tests with dichotomous items.

Akaike's Information Criterion (AIC)

To determine data dimensionality, it would be convenient to formulate a criterion to compare the likelihood of a *k*-factor model against that of the saturated model (Berger & Knol, 1990). Given Bock and Aitkin's (1981) ogive model, the probability of a correct response for ability vector $\vec{\theta}_j$ and item *i* is

$$P(X_{ij} = 1 \mid \vec{\theta}_j) = \Phi\left\{(\gamma_i - \sum_{k=1}^{m} \lambda_{ik}\theta_{jk}) / \sigma_i\right\}, \tag{5}$$

where γ_i is a threshold value for item *i*, θ_{jk} is the ability of person *j* on ability dimensional *k*, λ_{ik} is the loading of item *i* for dimension *k*. Akaike (1974) developed an information theoretic criterion for identifying the optimal and parsimonious models in data analysis. Akaike's information criterion is defined as:

$$AIC(m) = -2\ln\left[Lm\left(\hat{\theta}_j, \hat{\lambda}_i, \hat{\sigma}_i, \hat{\gamma}_i\right)\right] + 2K_m, \tag{6}$$

where $Lm\left(\hat{\theta}_j, \hat{\lambda}_i, \hat{\sigma}_i, \hat{\gamma}_i\right)$ is the maximized likelihood and K_m is the number of

independent parameters in the model. The term $2K_m$ is the penalty term which corrects for over-fitting due to increasing bias in the first term when the number of parameters in the model increases.

The term $AIC(m)$ is a measure of badness-of-fit, and the minimum value of the *AIC(m)* indicates the " true" dimensionality (Berger & Knol, 1990). The critical value of the *AIC* statistic is embodied in the penalty for over-fitting, and the Type I error rate decreases exponentially with increased sample size (McKinley, 1989). The *AIC* index has been recommended as a criterion for model selection, because when computed for a series of models of increasing dimensionality, it attains an optimum value for a model of intermediate dimensionality, thus allowing objective model selection (Berger & Knol, 1990; McDonald & Mok, 1995).

The practical performance of *AIC* in test data was not conclusive. Berger and Knol (1990) found that the *AIC* seemed to somewhat outperform the asymptotic χ^2 statistic, but these results were based on a small number of computer runs with sample sizes of 250 and 500. McKinley (1989) applied the *AIC* to artificial data fitting a confirmatory multidimensional item response model with the sample size of 1000, and found that *AIC* outperformed the likelihood ratio χ^2 test. McDonald (1989b) pointed out, however, that in applications, for a sufficiently small sample size, the optimum value must be attained by the unidimensional model, and for a sufficiently large sample size, it must be obtained by the saturated model. He concluded that *AIC* behaves just like the χ^2 significance test itself and cannot possibly be recommended for the use with real data.

Muthen's Robust Weighted Least Squares (Mplus)

Muthen proposed a probit function and a robust Weighted Least Squares (WLS) estimation procedure to assess dimensionality. This method was implemented in the computer program LISCOMP (B. Muthen, 1987) but later replaced by Mplus (L. K. Muthen & Muthen, 1998). According to Muthen (1978), the parameters of the factor analytic model for dichotomous variables can be estimated by minimizing the weighted least-square fit function

$$F = \frac{1}{2}(s-\sigma)'W^{-1}(s-\sigma), \tag{7}$$

where σ contains the population threshold and tetrachoric correlation values; s includes the sample estimates of the threshold and the sample tetrachoric correlation values; and W is a consistent estimator of the asymptotic covariance matrix of s, multiplied by the total sample size. The F function minimized in the WLS solution asymptotically follows a χ^2 distribution with $df=k(k-1)/2-t$, where k is the number of items and t is the number of parameters estimated in the model. If the null hypothesis in not true, the discrepancy function is distributed asymptotically as non-central chi-square. With WLS method, determining dimensionality is based on the fail-to-reject hypothesized model. That is, the hypothesis testing starts with the unidimensional model, and stopped when the hypothesized dimensionality is not rejected. In application, Stone and Yeh (2006) found that Mplus worked as well as NOHARM and TESTFACT when guessing was not modeled in the data. Tate (2003) also found that WLS procedure worked excellent for data with no guessing using an admittedly crude fit index equal to the ratio of χ^2 to degrees of freedom (χ^2/df). However, for data generated with guessing, this procedure generated distortions in the recovery of the true structure (Stone & Yeh, 2006).

2.2.3 Bivariate-Information Nonlinear Factor Analysis (NOHARM)

Starting from Spearman's common factor model, McDonald (1982) showed that IRT models are a special case of Nonlinear Factor Analysis (NLFA). He provided a general framework with a variety of models including unidimensional/multidimensional, linear/nonlinear, and dichotomous/polytomous models. The NOHARM program (Fraser, 1988) employs McDonald's (McDonald, 1981, 1982, 1985) NLFA, which uses a reparameterization of latent trait theory and " nonlinear harmonic" approximations to the normal ogive error distribution (Fraser & McDonald, 1988). In this process, the model is fit by unweighted least square which minimizes the squared difference between the observed frequencies of correctly answering item i and j, and the predicted frequencies of the joint occurrence of the pair of correct responses. Using McDonald's NLFA, researchers have developed various goodness-of-fit indices to decide the dimensionality of test data.

Approximate χ^2 Test of a Fitted NOHARM Model

Gessaroli and De Champlain (1996) proposed an approximate χ^2 test to assess dimensionality based on the estimation from NLFA. This approximate χ^2 statistic, originally proposed by Bartlett (1950) and outlined in Steiger (1980a; 1980b), tests whether all of the off-diagonal elements of the residual correlation matrix are equal to zero after fitting a k-factor NLFA model. The approximate χ^2 statistic is defined as

$$\chi^2 = (N-3)\sum_{i=1}^{k}\sum_{j=1}^{i=1} z_{ij}^{2(r)}, \qquad (8)$$

where $z_{ij}^{2(r)}$ is the square of Fisher's Z transformation corresponding to the residual correlation between item i and j, and N is the number of examinees in the sample. This statistic is distributed asymptotically as a χ^2 distribution with the degrees of freedom of $.5(m)\times(m-1)-t$, where m equals the number of items and t is the total number of independent parameters estimated in the NLFA model.

In an exploratory analysis based on adding successive factors to an initial unidimensional model, the search for an appropriate solution stops once the significance test indicates a good fit. Results from various simulation studies showed that this approximate χ^2 statistic is quite accurate in determining the number of factors underlying simulated item responses with small sample sizes (500 and 1000)(Gessaroli & De Champlain, 1996). However, Gessaroli and De Champlain (1996) also emphasized that this approximate χ^2 statistic has the same limitation as other χ^2 statistics: it tends to falsely reject the correct k-factor model with large sample sizes and fails to reject inaccurate models with small samples.

Residual Covariance Analyses after a Model Has Been Fitted to the Data

Based on the mathematical equivalence of the common IRT models and NLFA models, researchers (Choi, 1997; Hattie, 1984; McDonald, 1981, 1989a) suggested that a useful way to assess dimensionality is to analyze the residual covariance matrix obtained after fitting a model to an response matrix.

One of the model-data-fit indices reported in the NOHARM output, developed by Tanaka (1993), can be used with McDonald's NLFA model. This fit index is computed using $\gamma = 1 - Tr(R^2)/Tr(S^2)$, where R is the item residual covariance matrix, and S is

the matrix containing the raw product-moment of item pairs (Tanaka, 1993). A small value of γ implies that the residual covariances are close to the observed covariances, indicating a bad fit of the model. For practical application, Tanaka (1993) recommended that the value of γ should be greater than .95 for a model to be considered as good fit to the data. With this rule of thumb, McDonald & Mok (1995) used this index to assess the dimensionality of Law School Admission Test (LSAT) data and found that this index under-identified the second common factor.

Other residual covariance-based indices can be found in Berger and Knol (1990). Let Λ_k be the $n \times k$ estimated matrix of factor loadings from a solution with n items and k estimated common factors, and R be the tetrachoric correlation matrix. The off-diagonal elements of the residual matrix R^*, where $R^* = R - \Lambda_k \Lambda_k'$, are the residuals r_{ij}^*. Then, the equation of the mean squared residuals can be formulated as

$$f_1 = 2[n(n-1)]^{-1} \sum \sum_{i<j} (r_{ij}^*)^2 . \tag{9}$$

The mean absolute residuals is

$$f_2 = 2[n(n-1)]^{-1} \sum \sum_{i<j} | r_{ij}^* | . \tag{10}$$

As the formulas show, f_2 is less sensitive to outliers than f_1 because it uses the absolute value instead of the square in the equation, and thus is more often employed in previous studies. Hattie (1984; 1985) showed that f_2 can effectively discriminate between unidimensional and higher dimensional item response models after fitting the model by McDonald's NLFA. However, Hambleton and Rovinelli (1986) found that the residual analyses method provided disappointing results. The problem of applying this

criterion is its ambiguity of when the criterion is small enough to decide a good fit between the model and the data. In order to make a accurate decision based on f_2, a possible solution is to compare the criterion after the fit of a k-dimensional model with that from random data (Berger & Knol, 1990).

Another fit index using residuals after fitting the NLFA model is the Incremental Fit Index (IFI) proposed by De Champlain & Gessaroli (1991). The equation can be expressed as

$$IFI_k = \frac{SS_{reg}(k - factor) - SS_{reg}((k+1) - factor)}{SS_{reg}(k - factor)}. \qquad (11)$$

IFI calculates the proportion of the sum of squares of the residual covariances from the k-factor solution to that of the $(k+1)$-factor model. If the $(k+1)$-th factor is important in explaining the structure of the items, then the IFI should be quite large.

The theoretical advantage of these indices is that the assessment of dimensionality is made by an IRT-based model. The measure of model fit is directly related to the function minimized in the estimation procedure. However, there is an inherent weakness in this technique: there is no statistical significance test to decide the misfit of the model (De Champlain & Gessaroli, 1991).

2.2.4 Full-Information Item Factor Analysis (TESTFACT)

The computer program TESTFACT (Wilson et al., 2003) allows the practitioner to estimate the parameters and to fit various Full-Information Factor Analytic (FIFA) models. This estimation method uses the marginal maximum likelihood procedure outlined by Bock and Atkin (1981) via the expectation-maximization algorithm

(Dempster, Laird, & Rubin, 1977).

The FIFA uses information contained in the joint frequencies of the 2^n contingency tables of response counts on an n-item test. The probability of a correct response to an item is a function of an examinee's ability with respect to one or more latent factors and the location of the threshold parameters along the continuous variables. The thresholds and factor loadings are estimated so as to maximize the multidimensional probability function

$$L_m = P(X) = \frac{N!}{r_1!r_2!..r_s!} \widetilde{P}_i^{r1} \widetilde{P}_i^{r2} ... \widetilde{P}_i^{rs} \tag{12}$$

, where r_s is the frequency of response pattern s; and $\widetilde{P}_i$ is the marginal probability of the response pattern based on the item parameter estimates.

The user can assess the fit of a given FIFA model using a likelihood-ratio χ^2 statistic provided by TESTFACT. The FIFA yields a discrepancy function based on the ratio of the likelihood under the fitted model to the likelihood based on a saturated model, which fits the multidimensional distribution to the empirical frequencies. The likelihood-ratio χ^2 statistic can be defined as

$$G^2 = 2\sum_{l}^{2^n} r_l \ln(\frac{r_l}{N\widetilde{P}_l}), \tag{13}$$

where r_l is the frequency of response vector l, and $\widetilde{P}_l$ is the probability of response vector l. The degrees of freedom are $2^n - n(k+1) + k(k-1)/2$,where n is the number of items and k is the number of factors. The null hypothesis of this significance test is H_0: d= k. The decision about dimensionality is based on the point where the improvement of fit due to adding the net factor is not significant. If the G^2 is not significant, the

k-dimensional can be considered having good fit to the data. In this case, any additional factors could be attributed to sampling variation and therefore should not be interpreted. However, Mislevy (1986) found that this G^2 statistic often poorly approximates the χ^2 distribution given the large number of empty cells typically encountered with actual data sets. Moreover, Berger & Knol (1990) found that this G^2 test procedure erroneously favor the alternative hypothesis for almost all conditions.

Based on the work of Haberman (1977), equation (13) can be transformed to the likelihood-ratio G^2 difference test to assess the fit of a model. The statistic can be computed using the following expression

$$G^2_{diff} = G^2_{1F} - G^2_{2F}, \tag{14}$$

where G^2_{1F} is the value of the likelihood ratio G^2 statistic obtained after fitting a one-factor model, and G^2_{2F} is the likelihood ratio G^2 statistic from a two-factor model. The degrees of freedom are the difference between the *df* of one- and two-factor models. Again, the decision about dimensionality is made when the improvement of fit due to adding the proceeding factor is not significant. However, studies (Berger & Knol, 1990; De Champlain & Gessaroli, 1996) indicated that this likelihood-ratio G^2 difference test performs poorly for deciding the dimensionality of an item response matrix.

Overall, the goodness-of-fit indices proposed for the MIRT models in the literature can be summarized into two categories. The indices in the first category may tell the increase of fit or decrease of misfit when adding dimensions to the estimation model, but

k-dimensional can be considered having good fit to the data. In this case, any additional factors could be attributed to sampling variation and therefore should not be interpreted. However, Mislevy (1986) found that this G^2 statistic often poorly approximates the χ^2 distribution given the large number of empty cells typically encountered with actual data sets. Moreover, Berger & Knol (1990) found that this G^2 test procedure erroneously favor the alternative hypothesis for almost all conditions.

Based on the work of Haberman (1977), equation (13) can be transformed to the likelihood-ratio G^2 difference test to assess the fit of a model. The statistic can be computed using the following expression

$$G^2_{diff} = G^2_{1F} - G^2_{2F}, \tag{14}$$

where G^2_{1F} is the value of the likelihood ratio G^2 statistic obtained after fitting a one-factor model, and G^2_{2F} is the likelihood ratio G^2 statistic from a two-factor model. The degrees of freedom are the difference between the *df* of one- and two-factor models. Again, the decision about dimensionality is made when the improvement of fit due to adding the proceeding factor is not significant. However, studies (Berger & Knol, 1990; De Champlain & Gessaroli, 1996) indicated that this likelihood-ratio G^2 difference test performs poorly for deciding the dimensionality of an item response matrix.

Overall, the goodness-of-fit indices proposed for the MIRT models in the literature can be summarized into two categories. The indices in the first category may tell the increase of fit or decrease of misfit when adding dimensions to the estimation model, but

there is no significance test or justifiable criteria for deciding a good fit. The other category contains various χ^2 statistics. Even though a significance test is available for χ^2 statistics, the problem of deciding the data dimensionality is not yet solved. Kendall (1977) pointed out that Pearson's χ^2 and the likelihood ratio statistics are often regarded as equivalent because of their asymptotic properties. In practice, however, the large sample properties of χ^2 statistics are often unacceptable although the maximum likelihood estimators maintain standard large sample properties (Berger & Knol, 1990). What is more, the common problem for the χ^2 test for the fit of a model is its sensitivity to sample size. As McDonald (1989a) and Berger & Knol (1990) indicated, for large samples this procedure almost always rejects the null hypothesis and leads to wrong conclusions.

Therefore, in order to investigate the model identification problem, the ideal goodness-of-fit index should be able to reflect the degree of fit of the model to the data, and also not be overly sensitive to sample size. Besides this, the index should be reliable and also easy to interpret. To meet these requirements, a new index is introduced in the following sections.

2.3 The Development of Goodness-of-Fit Index for the MIRT Model

This research proposes a goodness-of-fit index applying the characteristic of R^2 to the MIRT model. In the first section, the basic relationship between R^2 and the likelihood ratio (LR) test is reviewed. Then, the likelihood-based R^2 analog proposed by Estrella (1998) for the dichotomous dependent variable (DDV) model is introduced. Lastly, the goodness-of-fit index based on the change of R^2 analog is proposed for describing the fit of a MIRT model.

2.3.1 The R^2 in the Ordinary Least Squares Model

Regression methods are an integral component of any data analysis concerning the relationship between a response variable and the explanatory variables (Hosmer & Lemeshow, 2000). The coefficient of determination, R^2, is a measure of how well the statistical model explains the observed data and is invariant to units of measurement. It describes the percentage of the total variance that can be explained by the regression model and becomes larger when the model fits the data better. The change of R^2 reflects the contribution of reducing residuals or improving overall model fit by adding an explanatory variable to the regression model. When it comes to select predictors for a multiple regression model, the change of R^2 is often used with the partial F test to decide if the inclusion of a predictor contributes significantly to the overall model fit.

Magee (1990) articulated the monotonic relationships between R^2 in the standard linear model, the Ward (W) statistic, and LR test statistic. On the basis of Magee's (1990) work, the inherent statistical characteristics of R^2 can be elucidated as follows:

Suppose that a dependent variable y has some functional relationship with the independent variable X,

$$y = \beta' X + \varepsilon, \tag{15}$$

where β is a set of parameters, and ε is the residual which consists of *iid* normal variates with a mean of 0. The first element of the β vector is generally considered as the intercept term, β_0. Let $\bar{y}$ denote the sample mean of y, and $\hat{y} = X(X'X)^{-1}X'y$, where $\hat{y}$ is the predicted value of y from the Ordinary Least Squares (OLS) model. The total sum of squares (SST) is $=(y-\bar{y})'(y-\bar{y})$, and the residual or error sum of squares (SSE)

is $(y-\hat{y})'(y-\hat{y})$. For the model containing only the intercept term β_0, $\hat{y}=\bar{y}$ and thus SSE is equal to the total variance SST. When an independent variable is added to the linear regression model, the decrease in SSE is due to the non-zero slope coefficient for that independent variable. To show the amount of error reduction by the independent variable, the R^2 statistic for the OLS model is defined as

$$R^2 = 1 - \frac{SSE}{SST}. \tag{16}$$

The term on the right-hand side indicates the percentage by which error is reduced. To test the null hypothesis, which means all the k-1 non-intercept elements of β are 0, the F test can be expressed as

$$F = \frac{(SST - SSE)/(k-1)}{SSE/(n-k)}. \tag{17}$$

And based on equation (16) and (17), it can be concluded that F is a monotonic increasing function of R^2 in the form of

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}. \tag{18}$$

Besides, if the error term in the OLS model is assumed to be normally distributed, F statistic is related to W statistic in the form of (e.g. Magee, 1990)

$$W = (k-1)\times\left(\frac{n}{n-k}\right)\times F, \tag{19}$$

given that $W = n\times\left(\frac{SST-SSE}{SSE}\right)$. Consequently, from equations (18) and (19), W can be reformulated as

$$W = n \times \left(\frac{R^2}{1 - R^2} \right). \tag{20}$$

In addition, Magee (1990) also showed that LR statistic for the same null hypothesis is

$$LR = -2\log\left(\frac{L_C}{L_U} \right) = -n \times \log(\frac{SSE}{SST}) = n \times \log(\frac{SST}{SSE}), \tag{21}$$

where $\log L_C = \text{constant} - \frac{n}{2}\log SST$ (log-likelihood of the fully constrained model)

$\log L_U = \text{constant} - \frac{n}{2}\log SSE$ (log-likelihood of an unconstrained model)

The model containing predictors is referred to as an unconstrained model because adding a predictor means relaxing a restriction in the maximization of the log-likelihood. Nagelkerke (1991) explained that the value of $-2\log L_C$ indicates the "error variation" of the model with only the intercept term. It is equivalent to the SST in the OLS model. With regard to the value of $-2\log L_U$, it is similar to the "error variation" for a model with predictors, analogous to the SSE in the OLS model (Menard, 2000). Under the null hypothesis that all the slopes in the population are 0, LR test follows a χ^2 distribution with k degree of freedom, where k is the number of predictors in the model.

In the standard linear model with normally distributed errors, there is a simple relationship between R^2 and LR statistic because LR is related to W (Vandaele, 1981) such that

$$LR = n \times \log\left(1 + \frac{W}{n} \right). \tag{22}$$

Form equation (20) and (22), the relationship between R^2 and LR can be formulated as

$$R^2 = 1 - \exp(\frac{-LR}{n}). \tag{23}$$

Just as R^2 in OLS model in equation (16) can be interpreted as the proportion of reduction in the error sum of squares, the likelihood-based R^2 in equation (23) can also be interpreted as the proportion of reduction in the -2log-likelihood statistic (Menard, 2000).

Moreover, Estrella (1998) demonstrated that the relationship between R^2 and LR statistic can also be expressed in terms of LR statistic per observation

$$A_{LR} = \frac{LR}{n} = -\frac{2}{n}\log\left(\frac{L_C}{L_U}\right), \tag{24}$$

which takes on values between 0 (misfit) and infinity (perfect fit). Accordingly to Estrella, equation (23) can be rewritten as

$$R^2 = 1 - (\frac{L_C}{L_U})^{\frac{2}{n}} = 1 - \exp(-A_{LR}). \tag{25}$$

The R^2 in equation (25) may be considered as a nonlinear rescaling of LR statistic per observation (Estrella, 1998). The endpoints of the scale are still compatible to a straightforward way indicating a "misfit" and a "perfect fit", respectively. Estrella (1998) also indicated that the difference in the likelihood statistic per observation is related to the difference in R^2 in an intuitive way such that

$$\frac{dR^2}{1 - R^2} = dA_{LR}. \tag{26}$$

The left side of this equation can be considered as a marginal R^2. This function specifies that the change of A_{LR} can be represented by the change of R^2. The marginal increment of fit, as shown to be consistent with the formal properties of R^2 in OLS, provides consistently accurate information to indicate goodness-of-fit (Estrella, 1998).

2.3.2 The R^2 Analog in the Dichotomous Dependent Variable Model

In the OLS model, the common assumption is that the error term of the model, ε, consists *iid* variates with a mean of zero and a fixed value of variance. This assumption is violated when the dependent variable in the regression model is dichotomously scored. In this case, a different regression model should be used for describing the relationship between the predictors and the dichotomized dependent variable. A Dichotomous Dependent Model (DDV) model can be defined in the form of a linear regression

$$y^* = \beta' x + \varepsilon, \tag{27}$$

where y^* is an unobservable variable, β is a vector of k coefficients (the first term is the intercept), and x is a vector of the values of k independent variables. In equation (27), y^* is linear in its parameters and may range from $-\infty$ to $+\infty$, depending on the range of x. There is also an observable variable y, which takes only two possible values and is related to y^* in the following way:

$y = 1$ if $y^* >$ threshold

$y = 0$, otherwise.

With dichotomous data, the outcome must be bounded between 0 and 1. The form of the estimation equation is $P(y = 1 \mid x) = F(\beta' x)$, where F is the cumulative distribution function of ε. In practice, F is usually specified as normal or logistic, but any other continuous distribution function whose first two derivatives exist and are well-behaved may be used (Estrella, 1998, p. 198). For a DDV model, the model parameters are estimated by maximum likelihood estimation, which can be defined as

$$L = \prod_{y_i=1} F(\beta' x_j) \prod_{y_i=0} [1 - F(\beta' x_j)]. \quad (28)$$

The likelihood function yields maximum likelihood estimators for the unknown parameters by maximizing the probability of obtaining the observed data. The resulting estimators are those that agree most closely with the observed data.

In the OLS model, there is only one reasonable residual variation criterion for the continuous dependent variable, but there are several possible variation criteria for DDV models (Efron, 1978). Based on the conceptual and mathematical similarity to the familiar R^2, many R^2analogies have been developed for the use with models having DDV (see Estrella, 1998; Kvalseth, 1985; Menard, 2000). In this study, the index proposed by Estrella (1998) was used to assess model-data-fit for test data because of its nice statistical properties. Estrella's measure of model-fit possesses the basic requirement of R^2 and has been used mainly in the areas of economics (Estrella, Rodrigues, & Schich, 2003; Herath & Takeya, 2003; Moneta, 2005; Shin & Moore, 2003; Stratmann, 2002) and medical research (Zheng & Agresti, 2000). Based on Esterlla's (1998) assertions, this goodness-of-fit index has some important statistical properties that other measures lack.

> This measure is constructed by imposing certain restrictions on its relationship with the underlying likelihood ratio statistics. These restrictions, including one expressed in terms of marginal increments in fit, are shown to be consistent with the formal properties of R^2 in the linear case and to provide consistently accurate signals as to statistical significance. This measure may be interpreted intuitively in a similar way to R^2 in the linear regression context, even away from the endpoints of its range values (Estrella, 1998, p. 198).

In the standard linear model with normally distributed errors, the relationship between R^2 and LR is clear. If there are n observations, of which n_1 indicates the case of

$$L = \prod_{y_i=1} F(\beta' x_j) \prod_{y_i=0} [1 - F(\beta' x_j)]. \quad (28)$$

The likelihood function yields maximum likelihood estimators for the unknown parameters by maximizing the probability of obtaining the observed data. The resulting estimators are those that agree most closely with the observed data.

In the OLS model, there is only one reasonable residual variation criterion for the continuous dependent variable, but there are several possible variation criteria for DDV models (Efron, 1978). Based on the conceptual and mathematical similarity to the familiar R^2, many R^2analogies have been developed for the use with models having DDV (see Estrella, 1998; Kvalseth, 1985; Menard, 2000). In this study, the index proposed by Estrella (1998) was used to assess model-data-fit for test data because of its nice statistical properties. Estrella's measure of model-fit possesses the basic requirement of R^2 and has been used mainly in the areas of economics (Estrella, Rodrigues, & Schich, 2003; Herath & Takeya, 2003; Moneta, 2005; Shin & Moore, 2003; Stratmann, 2002) and medical research (Zheng & Agresti, 2000). Based on Esterlla's (1998) assertions, this goodness-of-fit index has some important statistical properties that other measures lack.

> This measure is constructed by imposing certain restrictions on its relationship with the underlying likelihood ratio statistics. These restrictions, including one expressed in terms of marginal increments in fit, are shown to be consistent with the formal properties of R^2 in the linear case and to provide consistently accurate signals as to statistical significance. This measure may be interpreted intuitively in a similar way to R^2 in the linear regression context, even away from the endpoints of its range values (Estrella, 1998, p. 198).

In the standard linear model with normally distributed errors, the relationship between R^2 and LR is clear. If there are n observations, of which n_1 indicates the case of

$y = 1$. According to Estrella (1998), under the condition that H_0 is true (all the k-1 slopes are zero), equation (28) is maximized where $F(\beta_0) = \bar{y} = \frac{n_1}{n}$ and can be simplified as $L_C = \bar{y}^{n_1}(1-\bar{y})^{n-n_1}$ to represent the likelihood of the constrained model. Furthermore, he pointed out that the function of the log likelihood per observation has a particularly simple form that depends only on $\bar{y}$

$$A_C(\bar{y}) \equiv \frac{\ln L_C}{n} = \bar{y}\ln(\bar{y}) + (1-\bar{y})\ln(1-\bar{y}). \tag{29}$$

The hypothesis H_0 may be tested using LR statistic. When H_0 is true, the value of LR statistic is asymptotically distributed as a χ^2 with the degree of freedom of k-1.

With a dichotomous dependent variable, the approach using equation (25) fails because the LR statistic per observation is bounded (Estrella, 1998). Let A be the LR statistic per observation for DDV, then A can be expressed as

$$A = \frac{2}{n}\ln\left(\frac{L_U}{L_C}\right) = \frac{2}{n}(\ln L_U - \ln L_C). \tag{30}$$

When the model fits the data perfectly, the cumulative density function F can be represented as in Figure 2.3.1. In this case, when $L_U = 1$, A reaches its upper bound.

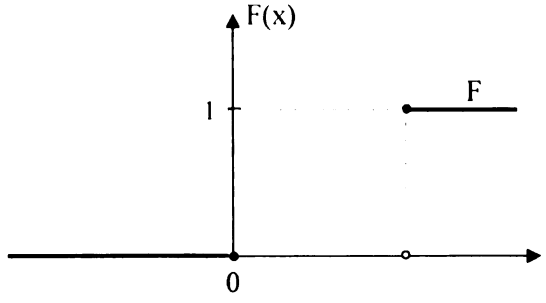


Figure 2.3.1. The cumulative density function F(x)

Estrella (1998) indicated that the upper bound of A can be expressed as $B=-\frac{2}{n}\ln L_{C} = -2A_{C}(\bar{y})$, where A_C is defined in equation (29). Based on this formula, the upper bound B is only a function of the log likelihood per observation. When $\bar{y}$ approaches either 0 or 1, B approaches 0.

The derivation of the R^2 analog is a differential equation, which bases primarily on an analog with the relationship between marginal R^2 and the Lagrange Multiplier (LM) statistic in the linear case (Estrella, 1998). The marginal R^2 in the linear case may be expressed in terms of the average LM statistic as (Estrella, 1998)

$$\frac{dR^2}{1-R^2} = \frac{dA_{LM}}{1-A_{LM}}. \tag{31}$$

The marginal R^2 increases with a rate inversely proportional to the distance between the current value of the statistic and its upper bound. In the DDV case, as Estrella (1998) explained, a measure based on the statistic A may be constructed using the fact that $0 \le A/B \le 1$. The index can be designed to reflect the marginal increase of fit being conversely proportional to $1-A/B$, which is the fraction of the "information content" of y that is still unexplained. The goodness-of-fit index, ϕ, can be defined by solving the differential equation (Estrella, 1998)

$$\frac{d\phi}{1-\phi} = \frac{dA}{(1-\frac{A}{B})}. \tag{32}$$

With the initial condition $\phi(0) = 0$, the solution of equation (32) is

$$\phi = 1-(1-\frac{A}{B})^{B} = 1-(\frac{\ln L_U}{\ln L_C})^{-\frac{2}{n}\ln L_C}. \tag{33}$$

To demonstrate the derivation of the fit index, the mathematical proof of equations (33) is shown in Appendix 1. When A=B, $\phi_0(B) = 1$, and this solution also satisfies the condition $\phi_0(B) = 1$ and $\phi_0'(0) = 1$ (Estrella, 1998). Moreover, Estrella (1998) pointed out that if B is replaced by "infinity" in the formula (33), then

$$\lim_{B \to \infty} 1 - (1 - A/B)^B = 1 - \exp(-A), \tag{34}$$

which is the exact expression for R^2 in the linear case in equation (25).

According to Estrella (1998), the goodness-of-fit index, ϕ, contains some desired features for a measure of model-data-fit. First, the measure takes on values on the unit interval and has the straightforward interpretation at the endpoints; that is, 0 corresponds to no fit and 1 corresponds to a perfect fit. The goodness-of-fit index is based on maximum likelihood method, which is also a common method used to calibrate test data in the field of educational measurement. This likelihood-based measure can be transformed into an F statistic as described in equation (18). Moreover, this index can work well for both the dichotomous and continuous dependent variables.

2.3.3 The *RLR* in the Multidimensional Item Response Model

Based on the similarity between the logistic regression model (one of the DDV models) and the logistic MIRT model (Reckase, 1985; Reckase & McKinley, 1991), it is possible to apply Estrella's R^2 analog to the MIRT model to reflect the error reduction by adding dimensions to the model. Furthermore, in order to reflect the degree of error reduction, the new index, which is the ratio of SSEs of two successive MIRT models, was proposed to show the improvement of model-data-fit.

If a DDV model takes the logistic function, it can be expressed as

$$y^* = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}, \tag{35}$$

where β_0 is the intercept parameter and β_1 indicates the vector of slope parameters. The observed variable y takes the value of 1 if y^* is greater than a threshold value and takes the value of 0 otherwise. The total number of model parameters needed to be estimated is expressed as $k+1$, where k is the number of predictors.

As indicated in Chapter 2.1, the logistic MIRT model is

$$P(U_{ij} = 1 | \vec{a}_i, d_i, \vec{\theta}_j) = \frac{\exp(d_i + \vec{a}_i' \vec{\theta}_j)}{1 + \exp(d_i + \vec{a}_i' \vec{\theta}_j)}, \tag{36}$$

where the $\vec{a}_i$, d_i, $\vec{\theta}_j$ are the same as those defined in Chapter 2.1. Compared to equation (35), d_i in equation (36) can be considered as the intercept parameter and the a_i vector can be viewed as the vector of slope parameters on the θ coordinate axes. The only difference between the two models is that the θ vector in equation (36) contains model parameters instead of predictors. In other words, along with the a and d parameters, the elements in the θ vector in the MIRT model also need to be estimated by the model. The total number of parameters in equation (36) is $n+f(n+m)$, where n is the number of items, f is the number of factors, and m is the number of examinees.

Employing the likelihood based R^2 analog to the MIRT model, the constrained MIRT model can be simplified as

$$P(U_{ij} = 1 | d_i) = \frac{\exp(d_i)}{1 + \exp(d_i)}. \tag{37}$$

This equation indicates that the probability of a correct response on item i depends only on d_i. Under this constrained model, d_i is estimated by n_1/n, where n_1 is the number of

examinees answering the item correctly, and n is the sample size. In this case, d_i in equation (37) can be considered as a nonlinear transformation of the item difficulty, also known as the p-value. Then, the probability of correctly answering an item only depends on the item difficulty and has nothing to do with the examinees' abilities. For the constrained model, the likelihood function can be expressed as

$$L_C = L(U \mid d_i) = \prod_{j=1}^{M} \prod_{i=1}^{n} P_i^{u_{ij}} (1 - P_i)^{1-u_{ij}}, \tag{38}$$

where u_{ij} takes on the value of 1 or 0, which indicates a correct or incorrect response respectively. The likelihood function for the unconstrained model (MIRT model) is

$$L_U = L(U \mid a_i, d_i, \theta_j) = \prod_{j=1}^{M} \prod_{i=1}^{n} P_{ij}^{u_{ij}} (1 - P_{ij})^{1-u_{ij}}, \tag{39}$$

where u_{ij} takes on the value of 1 or 0. The probability in equation (39) takes two subscripts representing a correct response of person j on item i. With Estrella's R^2 analog method, one can use the likelihood of the constrained model (L_C) and the likelihood of the unconstrained MIRT model (L_U) to express the proportion of the total variance explained by the MIRT model.

The feasibility of applying the R^2 analog to the MIRT model was first evaluated by examining the distribution of LR statistic. One of the well-known characteristics of the DDV model is that, when the null hypothesis (all the slopes in the model are 0 in the population) is true, LR statistic is χ^2 distributed. With the constrained model in equation (37), 1000 sets of item response data were generated for 25 items and 2000 examinees, and then were calibrated by the unidimensional MIRT model. The resulting distribution of LR statistic, as shown in Figure 2.3.1, has a mean of 38.47 and a variance of 70.605.

When taking sampling variation into account, this distribution approximates a χ^2 distribution since $\sigma^2 = 2\mu = 2\nu$, where ν are the degrees of freedom. This *LR* distribution demonstrates that the MIRT model contains the same characteristic as the DDV model, and thus can be considered as a special form of a DDV model.

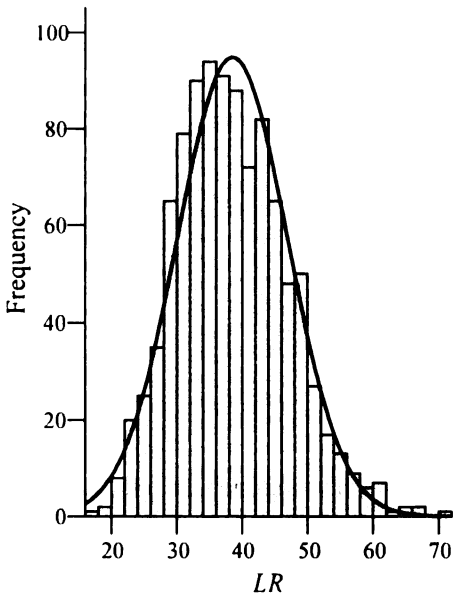


Figure 2.3.1. The observed distribution of *LR* statistic from the data generated by the constrained MIRT model

The R^2 analog can be used to represent how well the MIRT model fits the test data, but the most critical issue is to indicate whether or not the increase of fit by adding one more dimension to the model is important. In other words, it is useful to have an index reflecting the marginal effect of the "added" dimension to the overall model fit. Given a test data set, two successive MIRT models, the *k*-dimensional model and the *(k+1)*-dimensional model, are considered to describe the data. In order to indicate the

marginal effect of the $(k+1)$-th dimension to the overall model fit, the new index is defined as follows.

Let $\ln L_U^k$ be the log-likelihood of the k-dimensional MIRT model

$\ln L_U^{k+1}$ be the log-likelihood of the $(k+1)$-dimensional MIRT model

$\ln L_C$ be the log-likelihood of the constrained MIRT model

Then the R^2 analog for the two models can be expressed as

$$R_k^2 = 1 - \left(\frac{\ln L_U^k}{\ln L_C}\right)^{-\frac{2}{n}\ln L_C} \quad \text{and}$$

$$R_{k+1}^2 = 1 - \left(\frac{\ln L_U^{k+1}}{\ln L_C}\right)^{-\frac{2}{n}\ln L_C}.$$

Based on the equation (16), the percentage of the unexplained variance is $1 - R^2 = \frac{SSE}{SST}$. Taking the logarithm of both sides, the equation becomes $\ln(1 - R^2) = \ln(\frac{SSE}{SST})$. Then, the ratio of the log residuals (RLR) is defined as

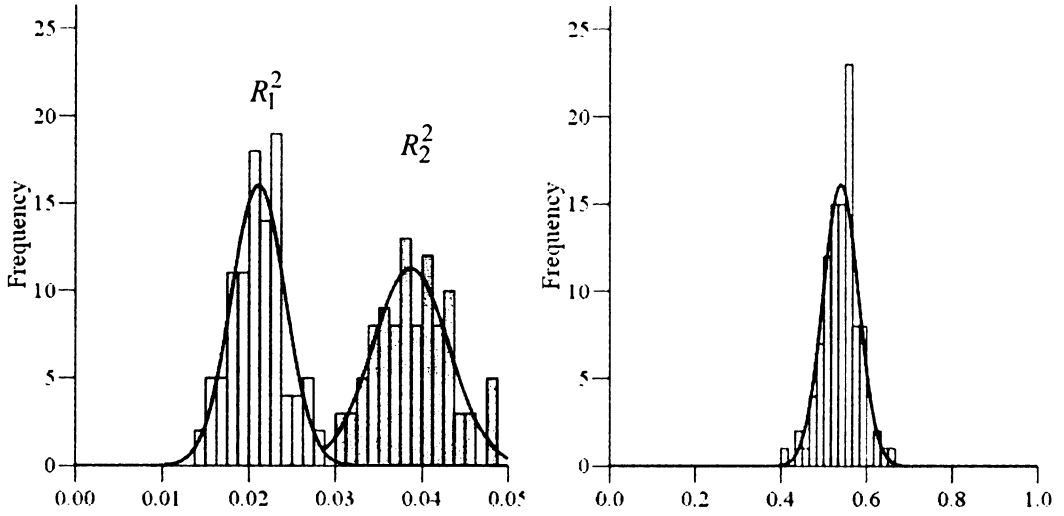
$$RLR = \frac{\ln(1 - R_k^2)}{\ln(1 - R_{k+1}^2)} = \frac{\ln(\frac{SSE_k}{SST})}{\ln(\frac{SSE_{k+1}}{SST})} = \frac{\ln(\frac{\ln L_U^k}{\ln L_C})}{\ln(\frac{\ln L_U^{k+1}}{\ln L_C})}. \qquad (40)$$

This index shows if the percentage of the unexplained variance in the $(k+1)$-th dimensional MIRT model is smaller than that in k-th dimensional MIRT model. The k-th dimension in equation (40) can be considered as the target dimension. The successive dimension, the $(k+1)$-th dimension, can be viewed as the reference dimension. Equation (40) focuses on the relative gain of overall model fit in view of comparing the

residuals in two models. If the k-dimensional model fits the data well, the reduction in SSE due to adding the $(k+1)$-th dimension should be minor. In this case, the value of the numerator and denominator in equation (40) are close to each other so that the *RLR* approaches 1. Since the *RLR* index always compares the SSEs for two successive models, for the convenience of discussion only the target dimension will be appended to the index to show the level of dimensionality. For instance, RLR_1 stands for the *RLR* index comparing the SSE of a one-factor model and that of a two-factor model.

The feasibility of using the R^2 analog and the *RLR* index to determine dimensionality is demonstrated by showing their empirical distributions in some basic cases. In all the following examples, 100 sets of item responses were generated for a 25-item test with 2000 examinees. For different situations, different models were used to generate the desired data.

When the data were generated by the constrained model, which only has the intercept term, no dimensionality underlies the data. When the data are explained by the MIRT model, the corresponding model-data-fit was reported in Figure 2.3.2. As Panel (A) shows, the distribution of R_1^2 has a mean of 0.0211 and a *SD* equal to 0.0031; the distribution of R_2^2 has a mean of 0.0387 and a *SD* of 0.0044. The small values of R_1^2 indicate that the unidimensional MIRT model explains little variance in the data. After adding the second dimension to the model, the value of R_2^2 has little increment, indicating limited increase in explained variance. The resulting distribution of RLR_1 has the distribution with the mean of 0.5391 and *SD* equal to 0.0412.

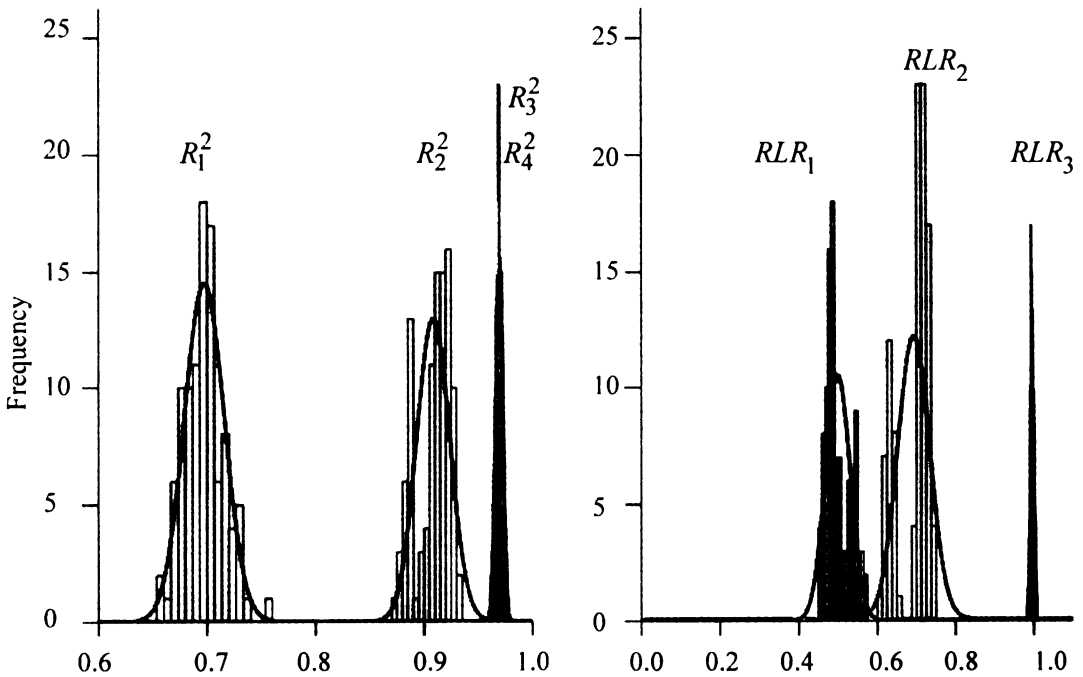


(A) The distributions of R_1^2 and R_2^2 (B) The distributions of RLR_1

Figure 2.3.2. The distributions of R_1^2, R_2^2, and RLR_1 for the constrained-model data

Another case offered here is the three-dimensional data. Item responses were generated assuming that three dimensions were independent of each other and all item discriminations equal to 1. As shown in panel (A) in Figure 2.3.3, the mean of R_1^2 is 0.6972 and the *SD* is 0.0183; the mean of R_2^2 is 0.9084 and the *SD* is 0.0183; the mean of R_3^2 is 0.9687 and the *SD* is 0.0033; the mean of R_4^2 is 0.97 and the *SD* is 0.003. Just like in the OLS model, the R^2 analog raises as the number of dimensions in the model increases. Regarding the distribution of R_3^2, when the model fits the data well, the index approaches 1. Besides, the distributions of RLR_3 and RLR_4 have substantial overlapping area, indicating the similarity of the two distributions. Thus, given that the model already fits the data well, the increase of fit by adding another dimension to the model is limited. Concerning the improvement of fit as shown in Panel (B); RLR_1 has a

mean of 0.4995 and a *SD* of 0.0315; RLR_2 has a mean of 0.6925 and a *SD* of 0.0410; RLR_3 has of mean .996 of and a *SD* of 0.004. When the model under-fits the data, the *RLR* is low and the distribution is located on the left side of the scale. Conversely, the index shifts to the right end of the scale with little variation when the model captures true dimensionality. The information from these distributions suggests that the *RLR* index offers clear and useful information about dimensionality.

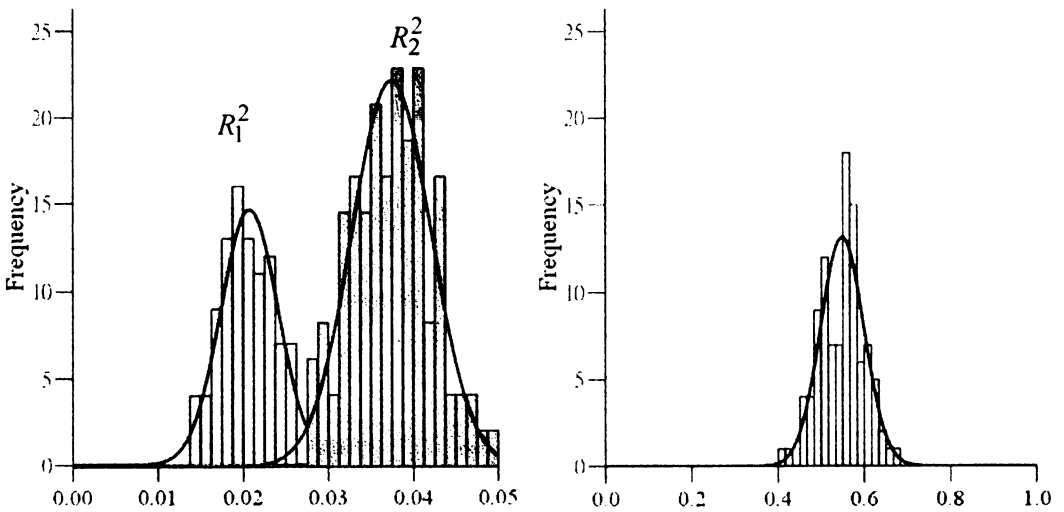


(A) The distributions of R_1^2, R_2^2, R_3^2, and R_4^2 (B) The distributions of RLR_1, RLR_2, and RLR_3

Figure 2.3.3. The distributions of R^2 and *RLR* for the three-dimensional data

An example of high-dimensional data was also offered to show the statistical characteristics of the proposed indices in the extreme situation. The item response data were generated with a 25-dimensional MIRT model assuming that all the dimensions

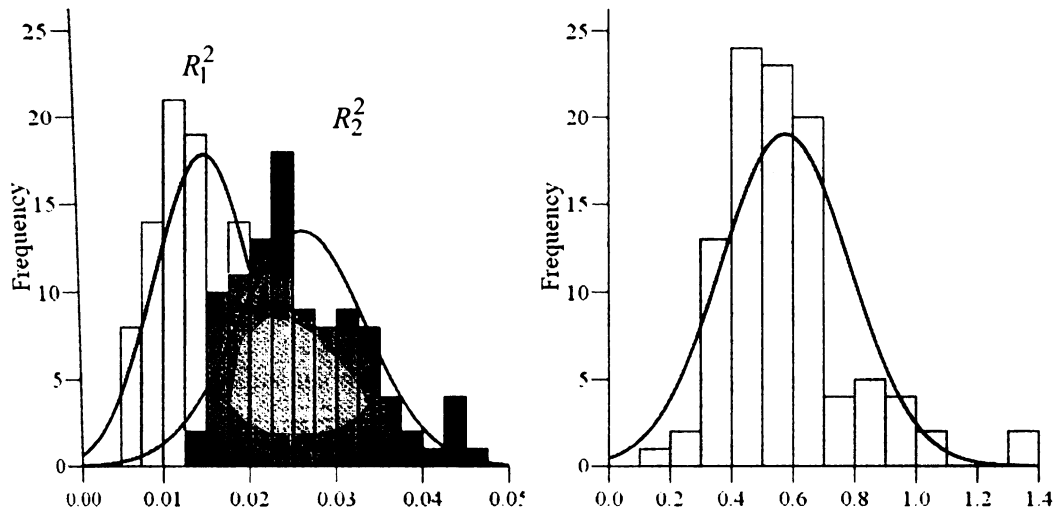
were independent of each other. Besides, the item discriminations were all fixed as 1.0. In this case, one item represented one distinct dimension in the data, and all the 25 dimensions had equal dominance of dimensionality. The results, as shown in Figure 2.3.4, indicated that the mean R_1^2 is 0.0208 and *SD* is 0.0034; the mean R_2^2 is 0.0374 and *SD* is 0.0047; RLR_1 has a distribution with mean 0.5487 and *SD* of 0.0505. The distributions of R_1^2, R_2^2, and RLR_1 are similar to those in the constrained model. The values of R_1^2 and R_2^2 indicate that the unidimensional and two-dimensional models only explain little variance in the data. These findings suggest that high dimensional data have similar properties as the constrained-model data. Because of the lack of a dominant factor, the increment of model-data-fit by adding dimensions to the model is limited. To explain the data well, complicated high-dimensional models need be employed.



(A) The distributions of R_1^2 and R_2^2 (B) The distribution of RLR_1

Figure 2.3.4. The distributions of R_1^2, R_2^2, and RLR_1 for the 25-dimensional model data

The last example offered here is to show how the R^2 analog and *RLR* index react to random data. For the distributions shown in Figure 2.3.5, R_1^2 has a mean of 0.0146 and a *SD* of 0.0056; R_2^2 has a mean of 0.0259 and a *SD* of 0.0074; RLR_1 has the mean of 0.5762 and a *SD* of 0.2098. Again, the means of R_1^2 and R_2^2 are as small as those in the constrained model and 25-dimensional model, but the variation is large. With random data, RLR_1 may have any value along the scale.



(A) The distributions of R_1^2 and R_2^2 (B) The distribution of RLR_1

Figure 2.3.5. The distributions of R_1^2, R_2^2, and RLR_1 for the random data

To summarize this chapter, there are several advantages of the *RLR* index as compared to other statistics.

(1) The calculation of *RLR* is based on maximum likelihood estimation, which is strong in its theoretical foundation, especially with a large sample size.

(2) This index has sound mathematical background. The derivation of the *RLR* index is based on the R^2 analog in the DDV model, which is in accordance with the R^2 in the linear regression model.

(3) *LR* statistics in the MIRT model is χ^2 distributed, which is consistent with the DDV model when the null hypothesis (all the slopes are zero) is true.

(4) With the *RLR* index, the dimensionality is assessed based on the improvement of the model-data-fit.

(5) The explanation of the *RLR* index is straightforward. The *RLR* index is viewed as the ratio of the log transformation of the unexplained percentage of the variance from two regression models. As shown in the preliminary simulations, the *RLR* index has a lower bound around .50. When the fit is good, the index approaches 1, indicating that the target dimension should be of use for describing the data.

(6) Furthermore, this statistic has the desirable property of showing the improvement of fit from adding dimensions to the model. Based on this procedure, researchers have a rule of thumb to decide when the increase of fit is important.

(7) Unlike the χ^2 test, the index is sensitive to sample size in a way that large sample size can increase the accuracy of identifying correct dimensionality. Within the limits of simulation, the index is not inflated by sample size and demonstrates desired statistical properties.

CHAPTER 3

METHOD

This chapter describes the research designs for exploring the statistical characteristics of the *RLR* index. Many researchers (Davey, Nering, & Thompson, 1997; Harwell, Stone, Hsu, & Kirisci, 1996) recommended the use of simulation studies because it offers an opportunity to permit theoretical results to be confirmed in practice. While manipulating all kinds of testing conditions, it is possible to know the statistical characteristics and the limits of the index of interest. With known dimensionality, two simulation studies representing some basic testing situations were conducted in order to explore the statistical properties of the *RLR* index. Furthermore, based on the procedures developed in simulation studies, the analysis of real test data is presented to demonstrate the feasibility of applying the fit index to a real testing situation.

3.1 Simulation Study I (Unidimensional Data Sets)

The focus of Study I is to explore the relationship between the *RLR* index and item characteristics for different unidimensional data. Correspondingly, the effects of test length and sample size on the *RLR* index are explored as well.

3.1.1 Research Design

Four variables were selected in Study I to simulate different testing conditions.

(1) Item discrimination (A)

When the MIRT model in equation (1) reduces to a unidimensional model, the value

of the *MDISC* is the same as the value of the *a*-parameter. In this study, the unidimensional data were generated in the unidimensional Rasch model fashion by setting all *a*-parameters equal in one test. The values of the *a*-parameters were fixed at four levels (0.2, 0.4, 0.6, and 0.8) with no variation in each data set, respectively. Low *a*-parameters imply that test items were poorly designed so that those items could not well differentiate examinees' abilities. Consequently, the signal in the test data may be weak and it would be difficult to identify the true dimensionality of the test data. High *a*-parameters indicate good items that can well differentiate examinees with different levels of ability. In this case, it is expected that the goodness-of-fit index can function well in recovering the true dimensionality.

Originally, the level of 1.0 of the *a*-parameter was included in the pilot study. When calibrated by multidimensional models, the simulation data with the *a*-parameters equal to 1.0 consistently generated a singular correlation matrix in TESTFACT. Because the calibrations for multidimensional models never succeeded, the level of 1.0 was excluded from Study I. This phenomenon implies that it is unlikely to have multidimensional solutions using full-information factor analysis when the item discriminations for unidimensional data are high. The procedure itself can detect the impossibility of getting multidimensional solution when the data are strongly unidimensional.

(2) Item Difficulty (*D*)

The variation in the distribution of item difficulty affects the sampling variability of tetrachoric correlations (Roznowski et al., 1991). When the spread of item difficulties increases, the tetrachoric correlation matrix tends to be non-Gramian and causes

computational difficulty in maximum likelihood factor analysis (McDonald, 1985). In order to explore how the variation of item difficulty affects full-information factor analysis and the *RLR* index, the *d*-parameters were sampled from normal distribution with a mean of 0 and three levels (0, 0.5, and 1) of standard deviation.

(3) Test Length (*T*)

To explore the possible effect of test length on the value of *RLR*, short test forms with 25 items and long test forms with 50 items were created. A short test was generated by selecting 25 *a*- and *d*-parameters from the predefined item distributions. With regard to a 50-item test, it was generated by adding parallel items to the original 25-item test. It is expected that as the number of items increases the data unidimensionality should be more accurately identified by the *RLR* index.

(4) Sample size (*S*)

According to the literature (Ackerman, 1994; R. L. Turner, Miller, Reckase, Davey, & Ackerman, 1996), usually 2000 or more examinees are suggested for MIRT calibration. In this study, the random samples of 2000 and 6000 examinees were drawn from a normal distribution with a mean of 0 and a standard deviation of 1. It is expected that the dimensionality index should vary in accuracy as a function of sample size.

3.1.2 Generation of Item Parameters and Response Patterns

Given the design of *a*-parameters (4), *d*-parameters (3), and test lengths (2), twenty-four combinations of simulated tests were generated. Table 3.1.1 tabulates the label and characteristics of each test. The numbers in the test label represent the levels of the *a*-parameters, *d*-parameters, and test length in order. Test 321, for example,

represents the test having the third level of the *a*-parameters (0.6), the second level of the *SD* of the *d*-parameters (0.5), and the first level of test length (25).

Table 3.1.1. Simulation tests for Study I

a-parameters	*SD* of *d*-parameters	short test form (25 items)	long test form (50 items)
0.2			
	0	Test 111	Test 112
	0.5	Test 121	Test 122
	1	Test 131	Test 132
0.4			
	0	Test 211	Test 212
	0.5	Test 221	Test 222
	1	Test 231	Test 232
0.6			
	0	Test 311	Test 312
	0.5	Test 321	Test 322
	1	Test 331	Test 332
0.8			
	0	Test 411	Test 412
	0.5	Test 421	Test 422
	1	Test 431	Test 432

When combining simulated tests (24) and sample sizes (2), forty-eight combinations of testing conditions were generated. In order to explore the consistency of the results in this study, replications are needed. For IRT-based studies, at least 25 replications have been recommended (Harwell et al., 1996). In this study, 100 sets of item response patterns were produced for each combination. Thus, the overall number of observations in Study I is 4800.

The way to generate dichotomous item response is to implement the known item parameters and ability parameters in the model in equation (1). Then, the computed probability is compared to a random number drawn from a uniform distribution ranged from 0 to 1. If the computed probability is greater than the random number, a response of 1 is generated, if not, a response of 0 is produced. The data simulation was

completed by using GENDAT5 developed by Thompson (Undated). This Fortran-based computer program uses input of the MIRT item parameters and an inter-factor correlation matrix, which is used to generate ability vectors based on the standardized normal distribution. This program can simulate multidimensional test data for up to 60 dimensions and can generate ability vector even for the case when factors are completely correlated in the correlation matrix.

3.1.3 Analysis Procedures and Computer Programs

The calculation of the *RLR* index depends upon being able to compute the maximum likelihood of the constrained model and that of the MIRT model. The likelihood of the constrained model was computed by the MATLAB program written by the author based on equation (38), and the likelihood of the MIRT model was calculated by TESTFACT (Wilson et al., 2003). Then, the values of the likelihood of the constrained model and the MIRT model were implemented in equation (40) to get the corresponding *RLR* value.

To decide data dimensionality, MIRT models with different levels of dimensionality were employed to analyze each data set. The test calibration started from the unidimensional MIRT model and continued to four-dimensional model. For each level of dimensionality the value of *RLR* was computed to reflect the increase of model-data-fit. After collecting the *RLR* values for all 4800 observations, the statistical package SPSS version 12.0 was employed to perform further statistical analyses. A Multidimensional Analysis of Variance (MANOVA) was conducted to explore the influence of the manipulated factors on the *RLR* index at different levels of dimensionality. Furthermore, the regression model was built to decide if the observed *RLR* index reflected a good fit

between the model and data.

3.1.4 Evaluation Criterion

The main purpose of Study I is to determine the level of accuracy of the *RLR* index in correctly determining unidimensionality. As shown in Figure 2.3.3, the distributions of the *RLR* index indicate that the *RLR* index is low and locates on the left side on the scale when the model under-fits the data; when the fit is good, the *RLR* index shifts to the right side of the scale and approaches 1. The theoretical conditional distribution of RLR_k can be expressed as Figure 4.1.1. When the null hypothesis is true (H_o: d= k), the distribution of RLR_k approaches 1 with small variation. Whenever the model under-fits the data, the

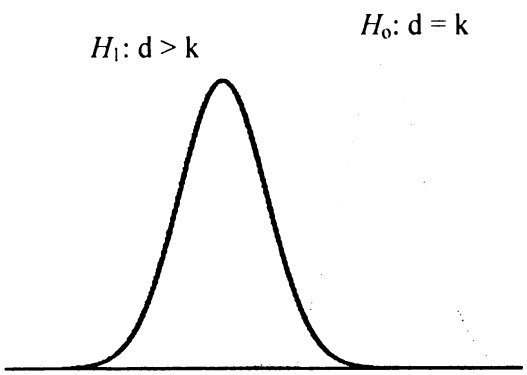


Figure 4.1.1 The theoretical distribution of RLR_k

In order to decide if a *RLR* value shows a good fit between the data and model, the 5% rejection criterion was set on the lower tail of the *RLR* distribution when the model

captures the true dimensionality. If the observed RLR_k is smaller than the lower bound of a good fit, the null hypothesis, H_o: d= k is true, is rejected. The significance test starts from testing the unidimensional model. If the observed RLR_1 index is less than the 5% lower bound, then the null hypothesis (H_o: d= 1) is rejected. Then the next significance test is to test if the observed RLR_2 index shows a good fit. Once a given value *RLR* is greater than the lower bound of a good fit, the null hypothesis is not rejected and the dimensionality can be decided.

To decide the lower bound of a good fit between the model and data, a regression analysis was conducted. Given the information of sample size, test length, the estimated a-parameters, and the estimated d-parameters, the predicted value of the *RLR* index can be estimated by the regression model. For each testing condition, the number of rejections obtained from the *RLR* index, and those from the G^2 test in equation (13) and the G^2_{diff} test in equation (14) were compared. The accuracy of these indices was deemed acceptable if the number of rejections in 100 replications was less than 5 for the true model. In Study I, it is expected that the *RLR* index should demonstrate lower Type I error rate than the G^2 test, and the G^2_{diff} test for the unidimensional data.

3.2 Simulation Study II (Multidimensional Data Sets)

The goal of the second simulation is to investigate how the *RLR* index detects dimensionality for different kinds of multidimensional test data. In this study, the two- and three-dimensional test data were generated under different conditions.

3.2.2 Research Design

The levels of multidimensionality were manipulated using three essential variables as follows:

(1) Inter-Factor Correlation (C)

In order to simulate examinees' multidimensional ability distributions, the correlation between factors (abilities) needs to be defined. The indices of dimensionality have long depended on relations among the successive eigenvalues obtained from factor analysis (see Hutten, 1980; Kaiser, 1970; Lord, 1980; Lumsden, 1957). The assumption of the scree test, for example, is that when the eigenvalues are displayed in their decreasing order, there will be a clear separation in fraction of total variance where the unimportant factor has been extracted. With information about the distribution of eigenvalues, Roznowski et al (1991) proposed the ratio difference index representing the ratio of the difference between the first two eigenvalues to their subsequent differences, in order to identify data unidimensionality. In this study, a different procedure was proposed. Dimensionality was manipulated by sampling correlation matrices in terms of the slope of eigenvalues and the determinant of the correlation matrix.

For a correlation matrix, the slope of eigenvalues reflects the magnitude and pattern of the inter-factor correlations. While working with the inter-factor correlations, the dimensional structure of the latent trait can be manipulated, and the level of dimensionality can be mapped on an arbitrary scale. An $n \times n$ correlation matrix M, for example, has n eigenvalues, $[\lambda_1 \quad \lambda_2 \quad \ldots \quad \lambda_n]$ that take the order $\lambda_1 \geq \lambda_2 .. \geq \lambda_n$. Given the same number of eigenvalues, when the distribution of eigenvalues is described by a

straight line, the slope of the straight line would indicate the relative importance of the underlying factors. Figure 3.2.1 is the scree plot showing the case of three 3×3 correlation matrices: $M_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$, $M_k = \begin{bmatrix} 1 & 0.8 & 0.6 \\ 0.8 & 1 & 0.4 \\ 0.6 & 0.4 & 1 \end{bmatrix}$, and $M_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

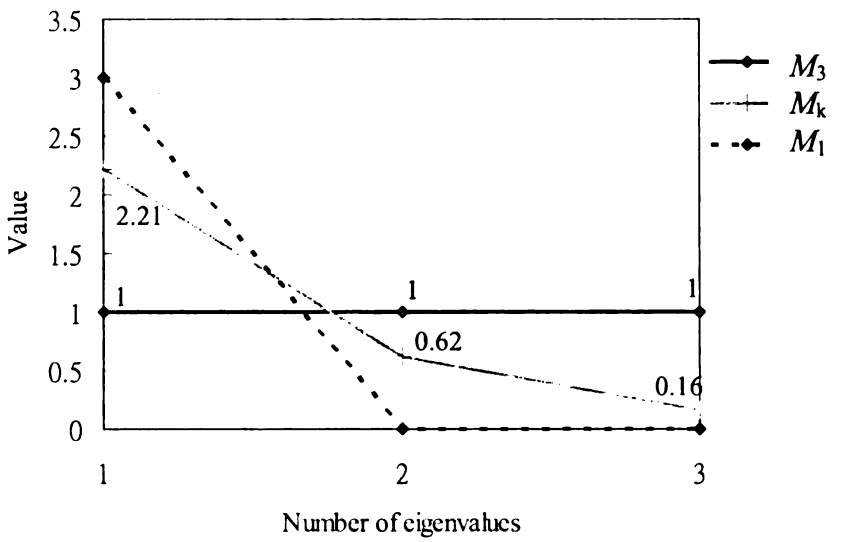


Figure 3.2.1. The scree plot of matrices M_1, M_k, and M_3

As shown in Figure 3.2.1, when the factors are completely independent of each other, such as the case of M_3, the eigenvalues of M_3 form a horizontal line so that the slope of the line is 0. The other extreme case occurred when the factors are completely dependent as shown in M_1. When the eigenvalues are fitted by a straight line, the slope is -1.5, which is the steepest slope among all possible 3×3 correlation matrices. It can be expected that when the inter-factor correlation is any number between 0 and 1, the slope of eigenvalues should fall in the interval between 0 (completely independent) and -1.5 (completely dependent). The correlation matrix M_k, for example, has the slope of

-1.02.

Furthermore, the determinant of the correlation matrix, det(M), has a functional relationship with its eigenvalues $[\lambda_1 \quad \lambda_2 \quad \ldots \quad \lambda_n]$, which is $\det(M) = \prod_{i=1}^{n} \lambda_i$. When factors are completely independent, as is the case for M_3, the determinant is 1; when factors are completely dependent, as is the case for M_1, than the determinant is 0. When the inter-factor correlations are not zero, the determinant of the correlation matrix should fall into the interval between 0 and 1. The correlation matrix M_k, for example, has the determinant as 0.2192.

For the correlation matrices of the same size, it is possible to differentiate different correlation matrices using the information of the slope of eigenvalues and the determinant of the correlation matrix. Figure 3.2.2 shows 3×3 correlation matrices with different levels of concentration of dimensionality represented by the slope of the eigenvalues and the determinant of the correlation matrix. The matrix M_3 has three factors that are completely independent of each other; the matrix, $M_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, represents a case when two of the three factors in the correlation matrix are completely dependent, but simultaneously, completely independent to the third factor. Since the rank of M_2 is two, the data with this correlation pattern can be considered as two-dimensional. Regarding the matrix M_1, since factors are completely correlated with each other, any data with this correlation pattern can be viewed as unidimensional.

The black dots in Figure 3.2.2 indicate the relationship between the determinant of the correlation matrix and the slope of eigenvalues when the inter-factor correlation was

manipulated by the design matrix $\begin{bmatrix} 1 & a & 0 \\ a & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, where $0 < a < 1$. When a equals 1, the design matrix becomes M_2; when a is 0, it becomes M_3. The trend of the black dots shows how the slope of eigenvalues and determinant varied when the three-dimensional data converged to two-dimensional data.

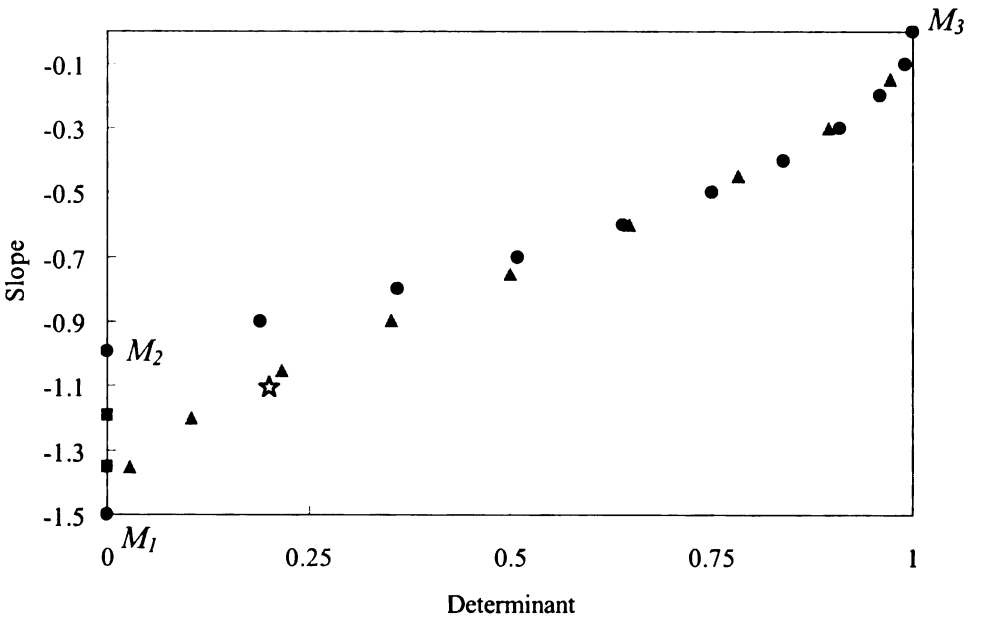


Figure 3.2.2. The relationship between the slope of eigenvalues and the determinant

The grey triangles, ▲, represent the relationship between determinant and slope of eigenvalues when the data converges from three dimension to one dimension with the design matrix $\begin{bmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{bmatrix}$, where $0 < a < 1$. When a equals 1, the design matrix becomes M_1; when a is 0, it becomes M_3. With regard to the grey squares, ■, they represent the case when two dimensions converge into one dimension by the design

matrix $\begin{bmatrix} 1 & 1 & a \\ 1 & 1 & a \\ a & a & 1 \end{bmatrix}$, where $0 < a < 1$. When a equals 1, the design matrix becomes M_1; when a is 0, it becomes M_2. Moreover, it is possible to locate a matrix whose off-diagonal elements are of any reasonable quantities for a correlation coefficient. The matrix M_k, for example, is located on Figure 3.2.2 with the star sign.

As shown in Figure 3.2.2, the relationship between the slope of eigenvalues and the determinant of the correlation matrix offers a way to summarize the concentration of dimensionality and also allows the comparison between correlation matrices. With this procedure, not only the degree of departure from unidimensionality but also the difference among different levels of multidimensionality can be laid out. In order to select the most representative correlation matrices for Study II, Figure 3.2.3 was created with grids specifying the space on the plane. As a result, six correlation matrices were selected:

$$C_1 = \begin{bmatrix} 1 & 1 & 0.7 \\ 1 & 1 & 0.7 \\ 0.7 & 0.7 & 1 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 1 & 1 & 0.4 \\ 1 & 1 & 0.4 \\ 0.4 & 0.4 & 1 \end{bmatrix}, \quad C_3 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad C_4 = \begin{bmatrix} 1 & 0.5 & 0.6 \\ 0.5 & 1 & 0.4 \\ 0.6 & 0.4 & 1 \end{bmatrix},$$

$$C_5 = \begin{bmatrix} 1 & 0.5 & 0.2 \\ 0.5 & 1 & 0.3 \\ 0.2 & 0.3 & 1 \end{bmatrix}, \text{ and } C_6 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

With these correlation matrices, the multidimensional abilities in Study II were generated from multivariate standard normal distribution.

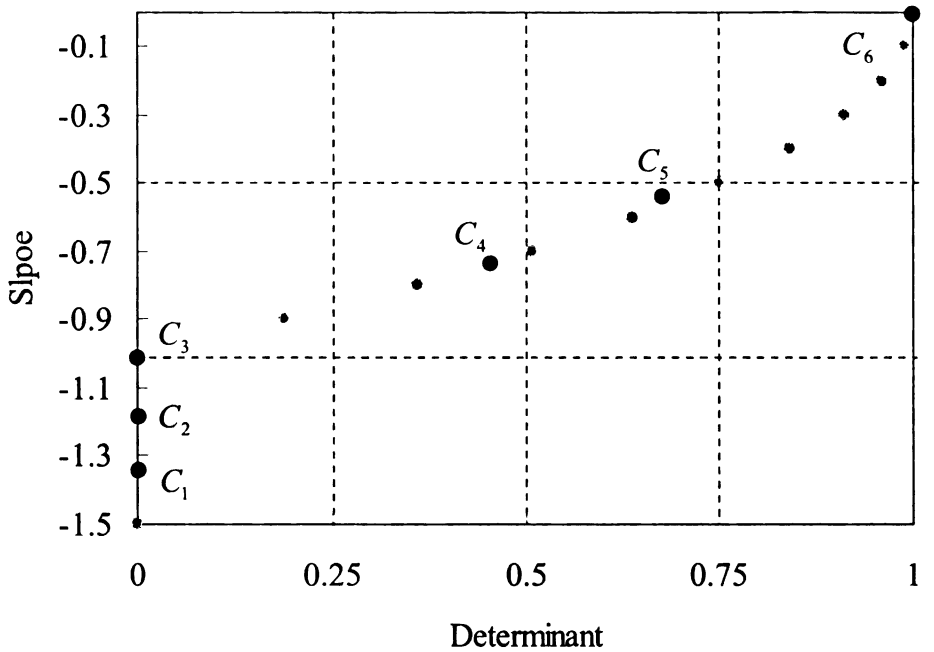


Figure 3.2.3. Selecting correlation matrices in terms of the slope of eigenvalues and the determinant of the correlation matrix

The simulations in Study II would be more complete if the correlation matrix $M_1=\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ were included. However, when including M_1 in this study, the large number of unsuccessful TESTFACT runs would generate a great number of missing observations for the data related to M_1 and cause problems in further statistical analysis. Thus, the matrix M_1 was not considered in Study II.

(2) Item-factor structure (*I*)

The simulation for multidimensional data were based on simple structure, which means that items have loading on one factor and zero loadings on the remaining factors. This type of item structure is desirable especially when evaluating scales created to measure either multiple constructs or components of a single construct (R. C. Turner,

2000).

Earlier studies (De Ayala & Hertzog, 1991; Gessaroli & De Champlain, 1996; Hambleton & Rovinelli, 1986) indicated that the number of items representing one factor was an important variable in simulating multidimensional test data. The item-factor structure indicates how well each factor was measured. When more items are sensitive to one factor, the data would have more information for that factor. Thus, it is anticipated that those factors can be easily identified by the statistical model. On the contrary, when a factor has only a few items, that dimension will be poorly measured. Accordingly, those factors may not be easily identified by the statistical model.

On the basis of the three-dimensional simple structure, the item-factor structure was manipulated by selecting different number of items to which each dimension related. The assignment of items to factors was listed in Table 3.2.1. Structure 1 shows the condition that the first 12 items measured factor 1, the second set of 12 items measured factor 2, and the remaining 24 items measured factor 3; Structure 2 represents the condition that the first sets of 16 items were indicators of factor 1, the second 16 items were indicators of factor 2, and the last 16 items were indicators of factor 3; Structure 3 shows the situation when the first 36 items related to factor 1, the second set of 6 items related to factor 2, and the last set of 6 items related to factor 3.

Table 3.2.1. Levels of the item-factor structure

Label	Number of items			Total
	Factor 1	Factor 2	Factor 3	
Structure 1	12	12	24	48
Structure 2	16	16	16	48
Structure 3	36	6	6	48

(3) Item discrimination (A)

Based on earlier studies on real tests, such as the ACT Mathematics Usage Test (Ackerman, 1994), LSAT (De Champlain & Gessaroli, 1996), TOEFL (McKinley & Way, 1992), and a nation-wide Math test for the 10 graders (R. L. Turner et al., 1996), the mean of *MDISC* often ranged from .76 to 1.34 and the *SD* varied from 0.2 to 0.5. In order to simulate item responses close to those from real tests, two levels of item discrimination were used in this study. The moderate level (M) of item discrimination was generated from $N(0.8\ ,\ 0.4^2)$; the high level (H) of item discrimination was generated from $N(1.2,\ 0.4^2)$.

As shown in Table 3.2.2, the research design in Study II generated thirty-six ($6\times3\times2$) combinations. Again, the levels for inter-factor correlation, item-factor structure, and item discrimination were labeled in order as the numbers in the form name. Form 321, for example, represents the test having the third level of the inter-factor correlation (C_3), the second level of the item-factor structure (16:16:16), and the first level of item discrimination (M).

Table 3.2.2. Simulated tests for Study II

Inter-factor correlation	Item discrimination	Item-factor structure		
		12:12:24	16:16:16	36:6:6
Two-dimension design				
$C_1 = \begin{bmatrix} 1 & 1 & 0.7 \\ 1 & 1 & 0.7 \\ 0.7 & 0.7 & 1 \end{bmatrix}$	M	Form 111	Form 121	Form 131
	H	Form 112 (50: 50)	Form 122 (67: 33)	Form 132 (88 : 12)
$C_2 = \begin{bmatrix} 1 & 1 & 0.4 \\ 1 & 1 & 0.4 \\ 0.4 & 0.4 & 1 \end{bmatrix}$	M	Form 211	Form 221	Form 231
	H	Form 212 (50: 50)	Form 222 (67: 33)	Form 232 (88 : 12)
$C_3 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	M	Form 311	Form 321	Form 331
	H	Form 312 (50: 50)	Form 322 (67: 33)	Form 332 (88 : 12)
Three-dimension design				
$C_4 = \begin{bmatrix} 1 & 0.5 & 0.6 \\ 0.5 & 1 & 0.4 \\ 0.6 & 0.4 & 1 \end{bmatrix}$	M	Form 411	Form 421	Form 431
	H	Form 412 (25: 25: 50)	Form 422 (33: 33: 33)	Form 432 (76: 12: 12)
$C_5 = \begin{bmatrix} 1 & 0.5 & 0.2 \\ 0.5 & 1 & 0.3 \\ 0.2 & 0.3 & 1 \end{bmatrix}$	M	Form 511	Form 521	Form 531
	H	Form 512 (25: 25: 50)	Form 522 (33: 33: 33)	Form 532 (76: 12: 12)
$C_6 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	M	Form 611	Form 621	Form 631
	H	Form 612 (25: 25: 50)	Form 622 (33: 33: 33)	Form 632 (76: 12: 12)

Under each test label, the numbers in the parentheses specified the percentage of items per dimension in the data. With the correlation matrices, C_1, C_2, and C_3, two-dimensional data were generated because the first two factors converged into one factor. Thus, those items originally sensitive to the first and second factors would converge into a bigger item cluster. With structure 1, 50% of the items loaded on the converged first dimension, and the remaining 50% of items loaded on the other

dimension. With regard to structure 2, 67% of the items were grouped as the first dimension and the rest of the 33% items grouped as the second dimension. With respect to structure 3, 88% of the items were clustered as one dimension and the remaining 12% formed a second dimension. Regarding the correlation matrices C_4, C_5, and C_6, three-dimensional data were generated and the percentage of item per dimension was consistent with the original item-factor structure.

3.2.2 Generation of Item Response Patterns

The d-parameters were randomly generated from a normal distribution $N(0, 1)$ for all 48 test items. The multidimensional ability distributions were generated from the standardized multidimensional normal distribution with the pre-selected inter-factor correlation matrices. Again, the sample size used in Study II was 2000. The procedures for generating item response patterns were the same as those described in Section 3.1.2. For each cell of the thirty-six combinations, 100 replications were performed, and the total number of 3600 multidimensional data sets was produced.

3.2.3 Procedures and Computer programs

The procedures for computing of the *RLR* index were the same as those described in section 3.1.3. In study II, the test calibration started from the unidimensional model and continued to the 5-dimensional model. For each level of dimensionality, the *RLR* index was computed to show the improvement of model-data-fit.

3.2.4 Evaluation Criterion

Again, the statistical properties of the *RLR* index were explored and compared with those of the G^2 test and the G^2_{diff} test. To test whether the data can be well fit by the unidimensional model, the unidimensional regression model generated in Study I was used in conjunction with sample size, test length, and the estimated unidimensional item parameters. If the observed RLR_1 is smaller than the predicted lower bound, then the null hypothesis (H_0: d=1) was rejected, indicating a higher-dimensional model is needed.

To test whether or not the null hypothesis (H_0: d=2) was true for a given data sets, the two-dimensional regression model was constructed based on the two-dimensional data. Again, given that the model captures the true two-dimensional data, the regression model sets up the 5% rejection area at the lower end of the predicted RLR_2 distribution. If the observed RLR_2 value is smaller than the predicted lower bound, the null hypothesis is rejected and the data should be modeled with higher dimension. Using the same procedure, the three-dimensional regression model was constructed to test the null hypothesis (H_0: d=3) based on the three-dimensional model. If the observed RLR_3 value is smaller than the predicted lower bound, then the data should be modeled with higher dimension. It is expected that, the number of false rejections should be lower than 5 among 100 replications when the regression model captures the true dimensionality. Conversely, when detecting the wrong models, the *RLR* index should generate large number of rejections, indicating high statistical power.

3.3 Real Data Analysis

Along with the simulation studies, the statewide test data of the Mathematics Test from the Michigan Educational Assessment Progress (MEAP) testing program were

analyzed. Under the No Child Left Behind (NCLB) act of 2001, the federal approval depends on strict alignment of state assessment to state content standards. Michigan's Mathematics Test, which developed to match the mathematics content standards, were developed to measure what Michigan educators believe all students should learn and be able to achieve in each grade level (Michigan Department of Education, 2004).

In this study, the test data from the Grade 4 Mathematics Test were used. The Mathematics Test contained 57 items covering content knowledge in data and probability, geometry, measurement, and numbers and operations. To be more precise, students were requested to demonstrate their academic proficiency in (1) fluency with operations and estimations; (2) geometric shape, properties, and mathematical arguments; (3) meaning, notation, place value, and comparisons; (4) number relationships and meaning of operations; (5) problem solving involving measurement; (6) data representation; (7) spatial reasoning and geometric modeling; and (8) units and systems of measurement (Michigan Department of Education, 2006). Students who score high on the test have documented substantial achievement in mathematics at the grade-4 level. In terms of the hierarchical ability structure in the blueprint of the Mathematics Test, it is suspected that the resulting test data may be explained by a multidimensional model.

The test data from 10000 examinees were requested from the testing program. The sample was then divided into five smaller data sets with 2000 examinees by random selection. The MIRT model parameters for different levels of dimensionality were estimated using TESTFACT. For each level of dimensionality, the corresponding *RLR* index was computed to determine the increment of model-data-fit. To decide the

dimensionality of MEAP data, the regression models developed from the simulation studies were used to determine whether the observed *RLR* index showed a good fit between the model and data. If the observed *RLR* index fell in the 5% rejection area of the lower end, the null hypothesis was rejected, and the higher-dimensional model was tested in turn. The significance test started from the unidimensional model and stopped when the null hypothesis was not rejected. Instead of making judgments form a single test, the results from different sample data would give the basis of cross-validation and offer a more dependable decision.

CHAPTER 4

RESULTS

Based on the research designs described in the previous chapter, the main results of the three studies are provided along with the initial interpretations.

4.1 Simulation Study I (Unidimensional Data Sets)

The focus of Study I was to explore the effects of item discrimination (A), item difficulty (D), sample size (S), and test length (T) on the *RLR* index. However, when the unidimensional data were analyzed by multidimensional models, some of the TESTFACT analyses failed. When T was short (25 items), all TESTFACT runs were successful regardless of the levels of A, D, and S. When T was long (50 items), some tests generated a singular tetrachoric correlation matrix, causing a serious estimation problem in full-information factor analysis. Table 4.1.1 shows the number of unsuccessful cases out of 100 replications for long-test data. Given that T was long, when D was high, the probability of getting a singular tetrachoric correlation matrix was high, especially for the case when S was small (2000). For these data sets, the rates of getting a singular tetrachoric correlation matrix increased with the increment of the number of factors in the estimation model. The highest rate of getting unsuccessful TESTFACT runs occurred when the unidimensional data were analyzed by the four-dimensional MIRT model.

Table 4.1.1. The number of unsuccessful TESTFACT runs for long tests in Study I

Sample size	Test	MIRT Model			
		1 Factor	2 Factor	3 Factor	4 Factor
2000					
	Test 112	0	0	0	0
	Test 122	0	0	0	2
	Test 132	0	0	4	15
	Test 212	0	0	0	0
	Test 222	0	0	0	0
	Test 232	0	1	3	35
	Test 312	0	0	0	0
	Test 322	0	0	0	0
	Test 332	0	0	3	18
	Test 412	0	0	0	0
	Test 422	0	0	0	0
	Test 432	0	0	2	7
6000					
	Test 112	0	0	0	0
	Test 122	0	0	0	0
	Test 132	0	0	0	4
	Test 212	0	0	0	0
	Test 222	0	0	0	0
	Test 232	0	0	3	7
	Test 312	0	0	0	0
	Test 322	0	0	0	0
	Test 332	0	0	0	1
	Test 412	0	0	0	0
	Test 422	0	0	0	0
	Test 432	0	0	0	0

Note: The results for short tests were not listed because all TESTFACT runs were successful.

4.1.1 Results of the Summary Statistics

With regard to those successful TESTFACT runs, no outliers were found in the preliminary analysis. Table 4.1.2 and Table 4.1.3 display the summary statistics of the *RLR* values in each condition. The changes of *RLR* values associated with dimensionality were plotted in Figure 4.1.1 to Figure 4.1.4. The conditional

distributions of *RLR* values were presented in Appendix B as a supplement to the summary statistics.

By and large, the *SD* of the *RLR* values in each condition was small. Given the same levels of *S* and *T*, the *SD* of the *RLR* values was small when *A* was high. Conditioned on *A* and *D*, the *SD* of the *RLR* values decreased when *T* was long or *S* was large. For most data sets, the *SD* of the *RLR* values for a higher-factor model was smaller than that for a lower-factor model. The decrease of the variation of the *RLR* values was more noticeable when *A* was low.

The *RLR* index for the unidimensional model was particularly sensitive to item parameters. The increase of *A* was proportional to RLR_1, but the increase of *D* was inversely proportional to RLR_1. The effects of *A* and *D* on RLR_1 was similar across different combinations of *S* and *T*.

When the *RLR* values were plotted against dimensionality, the lines indicated the change of the *RLR* values as a result of dimensionality. As shown from Figure 4.1.1 to Figure 4.1.4, the color of the lines denotes different levels of *A*, and the shape of the lines represents different levels of *D*. For the tests with *A* higher than 0.2, the *RLR* values were all centered to 1 and formed horizontal lines. The change of the *RLR* values was limited when adding more factors to the model. Since the increase of the *RLR* values due to adding factors to the model was trivial, this pattern of the *RLR* values might imply that the unidimensional model was good enough to explain the test data. Conversely, for the tests with *A* equal to 0.2, the *RLR* values showed noticeable increase associated with dimensionality, especially when *D* was large, *S* was small, and *T* was short. This pattern implied that higher-factor models fit the data better than the unidimensional model.

Table 4.1.2. Summary statistics of the *RLR* index for short tests

25-item test	*RLR*	2000 examinees				6000 examinees			
		Mean	*SD*	*N*	*SE*	Mean	*SD*	*N*	*SE*
Test 111									
	RLR_1	0.8713	0.0224	100	0.0022	0.9506	0.0085	100	0.0008
	RLR_2	0.9046	0.0156	100	0.0016	0.9614	0.0059	100	0.0006
	RLR_3	0.9225	0.0110	100	0.0011	0.9679	0.0042	100	0.0004
Test 121									
	RLR_1	0.8533	0.0254	100	0.0025	0.9429	0.0093	100	0.0009
	RLR_2	0.8942	0.0171	100	0.0017	0.9542	0.0080	100	0.0008
	RLR_3	0.9152	0.0143	100	0.0014	0.9639	0.0061	100	0.0006
Test 131									
	RLR_1	0.8086	0.0356	100	0.0036	0.9245	0.0143	100	0.0014
	RLR_2	0.8695	0.0231	100	0.0023	0.9398	0.0115	100	0.0011
	RLR_3	0.8933	0.0182	100	0.0018	0.9508	0.0099	100	0.0010
Test 211									
	RLR_1	0.9809	0.0034	100	0.0003	0.9935	0.0012	100	0.0001
	RLR_2	0.9843	0.0024	100	0.0002	0.9947	0.0008	100	0.0001
	RLR_3	0.9862	0.0020	100	0.0002	0.9954	0.0008	100	0.0001
Test 221									
	RLR_1	0.9783	0.0039	100	0.0004	0.9925	0.0012	100	0.0001
	RLR_2	0.9823	0.0029	100	0.0003	0.9940	0.0010	100	0.0001
	RLR_3	0.9844	0.0023	100	0.0002	0.9949	0.0009	100	0.0001
Test 231									
	RLR_1	0.9717	0.0050	100	0.0005	0.9901	0.0017	100	0.0002
	RLR_2	0.9771	0.0042	100	0.0004	0.9921	0.0018	100	0.0002
	RLR_3	0.9791	0.0039	100	0.0004	0.9930	0.0016	100	0.0002
Test 311									
	RLR_1	0.9924	0.0011	100	0.0001	0.9975	0.0005	100	0.0000
	RLR_2	0.9937	0.0009	100	0.0001	0.9979	0.0003	100	0.0000
	RLR_3	0.9944	0.0009	100	0.0001	0.9984	0.0003	100	0.0000
Test 321									
	RLR_1	0.9917	0.0012	100	0.0001	0.9972	0.0005	100	0.0000
	RLR_2	0.9932	0.0009	100	0.0001	0.9977	0.0003	100	0.0000
	RLR_3	0.9939	0.0011	100	0.0001	0.9982	0.0003	100	0.0000
Test 331									
	RLR_1	0.9898	0.0018	100	0.0002	0.9966	0.0005	100	0.0001
	RLR_2	0.9915	0.0014	100	0.0001	0.9971	0.0006	100	0.0001
	RLR_3	0.9920	0.0017	100	0.0002	0.9975	0.0006	100	0.0001
Test 411									
	RLR_1	0.9955	0.0007	100	0.0001	0.9984	0.0003	100	0.0000
	RLR_2	0.9963	0.0006	100	0.0001	0.9990	0.0002	100	0.0000
	RLR_3	0.9969	0.0006	100	0.0001	0.9994	0.0003	100	0.0000
Test 421									
	RLR_1	0.9952	0.0008	100	0.0001	0.9984	0.0003	100	0.0000
	RLR_2	0.9960	0.0006	100	0.0001	0.9988	0.0002	100	0.0000
	RLR_3	0.9967	0.0008	100	0.0001	0.9993	0.0003	100	0.0000
Test 431									
	RLR_1	0.9942	0.0010	100	0.0001	0.9981	0.0003	100	0.0000
	RLR_2	0.9951	0.0008	100	0.0001	0.9984	0.0003	100	0.0000
	RLR_3	0.9956	0.0009	100	0.0001	0.9989	0.0003	100	0.0000

Table 4.1.3. Summary statistics of the *RLR* index for long tests

50-item test	*RLR*	2000 examinees				6000 examinees			
		Mean	*SD*	*N*	*SE*	Mean	*SD*	*N*	*SE*
Test 112									
	RLR_1	0.9096	0.0117	100	0.0012	0.9673	0.0039	100	0.0004
	RLR_2	0.9257	0.0087	100	0.0009	0.9718	0.0027	100	0.0003
	RLR_3	0.9353	0.0063	100	0.0006	0.9750	0.0024	100	0.0002
Test 122									
	RLR_1	0.8982	0.0127	100	0.0013	0.9623	0.0044	100	0.0004
	RLR_2	0.9177	0.0087	100	0.0009	0.9668	0.0037	100	0.0004
	RLR_3	0.9270	0.0070	98	0.0007	0.9707	0.0027	100	0.0003
Test 132									
	RLR_1	0.8766	0.0159	100	0.0016	0.9536	0.0051	100	0.0005
	RLR_2	0.9004	0.0114	96	0.0012	0.9595	0.0044	100	0.0004
	RLR_3	0.9133	0.0097	83	0.0011	0.9639	0.0046	96	0.0005
Test 212									
	RLR_1	0.9844	0.0019	100	0.0002	0.9948	0.0006	100	0.0001
	RLR_2	0.9867	0.0012	100	0.0001	0.9954	0.0004	100	0.0000
	RLR_3	0.9871	0.0011	100	0.0001	0.9957	0.0004	100	0.0000
Test 222									
	RLR_1	0.9827	0.0020	100	0.0002	0.9941	0.0007	100	0.0001
	RLR_2	0.9848	0.0017	100	0.0002	0.9948	0.0006	100	0.0001
	RLR_3	0.9857	0.0015	100	0.0001	0.9952	0.0005	100	0.0001
Test 232									
	RLR_1	0.9793	0.0025	99	0.0003	0.9929	0.0009	100	0.0001
	RLR_2	0.9817	0.0019	97	0.0002	0.9939	0.0007	97	0.0001
	RLR_3	0.9828	0.0029	63	0.0004	0.9942	0.0007	90	0.0001
Test 312									
	RLR_1	0.9931	0.0007	100	0.0001	0.9976	0.0003	100	0.0000
	RLR_2	0.9942	0.0007	100	0.0001	0.9983	0.0002	100	0.0000
	RLR_3	0.9943	0.0007	100	0.0001	0.9983	0.0003	100	0.0000
Test 322									
	RLR_1	0.9925	0.0008	100	0.0001	0.9974	0.0003	100	0.0000
	RLR_2	0.9936	0.0007	100	0.0001	0.9980	0.0003	100	0.0000
	RLR_3	0.9936	0.0008	100	0.0001	0.9981	0.0004	100	0.0000
Test 332									
	RLR_1	0.9914	0.0010	100	0.0001	0.9971	0.0003	100	0.0000
	RLR_2	0.9925	0.0008	97	0.0001	0.9976	0.0003	100	0.0000
	RLR_3	0.9934	0.0008	77	0.0001	0.9978	0.0004	99	0.0000
Test 412									
	RLR_1	0.9946	0.0006	100	0.0001	0.9973	0.0003	100	0.0000
	RLR_2	0.9963	0.0005	100	0.0001	0.9987	0.0002	100	0.0000
	RLR_3	0.9968	0.0006	100	0.0001	0.9994	0.0003	100	0.0000
Test 422									
	RLR_1	0.9945	0.0006	100	0.0001	0.9975	0.0002	100	0.0000
	RLR_2	0.9959	0.0005	100	0.0001	0.9987	0.0002	100	0.0000
	RLR_3	0.9965	0.0007	100	0.0001	0.9993	0.0002	100	0.0000
Test 432									
	RLR_1	0.9942	0.0007	100	0.0001	0.9975	0.0003	100	0.0000
	RLR_2	0.9954	0.0006	98	0.0001	0.9985	0.0002	100	0.0000
	RLR_3	0.9959	0.0007	92	0.0001	0.9991	0.0003	100	0.0000

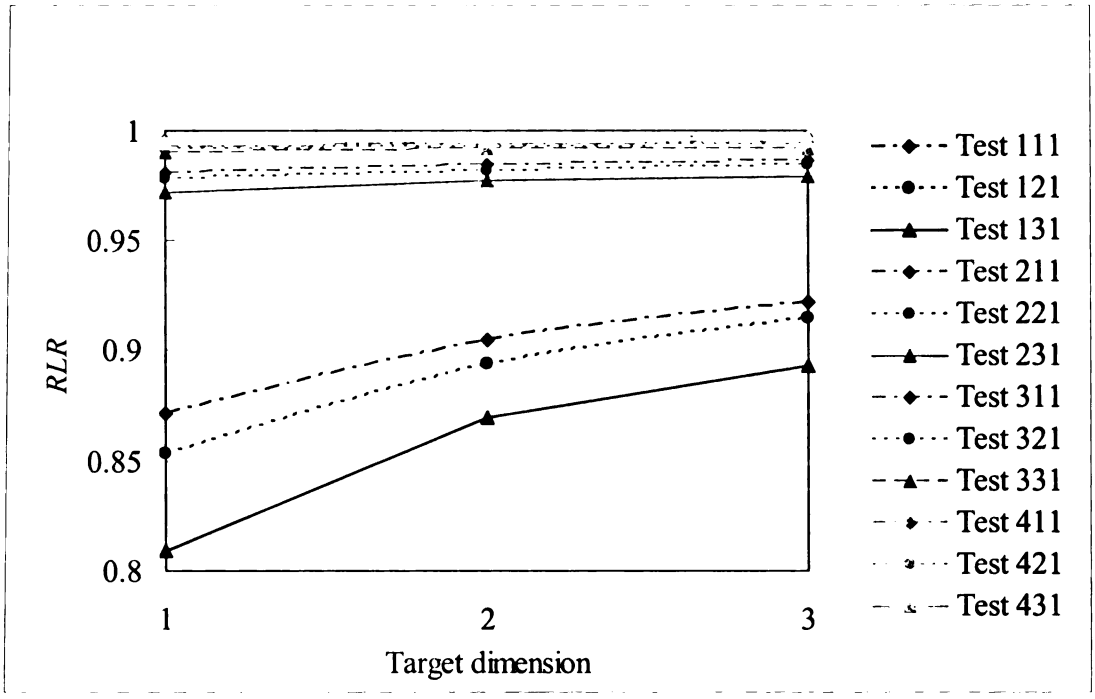


Figure 4.1.1. The change of *RLR* with dimensionality for a 25-item test and 2000 examinees

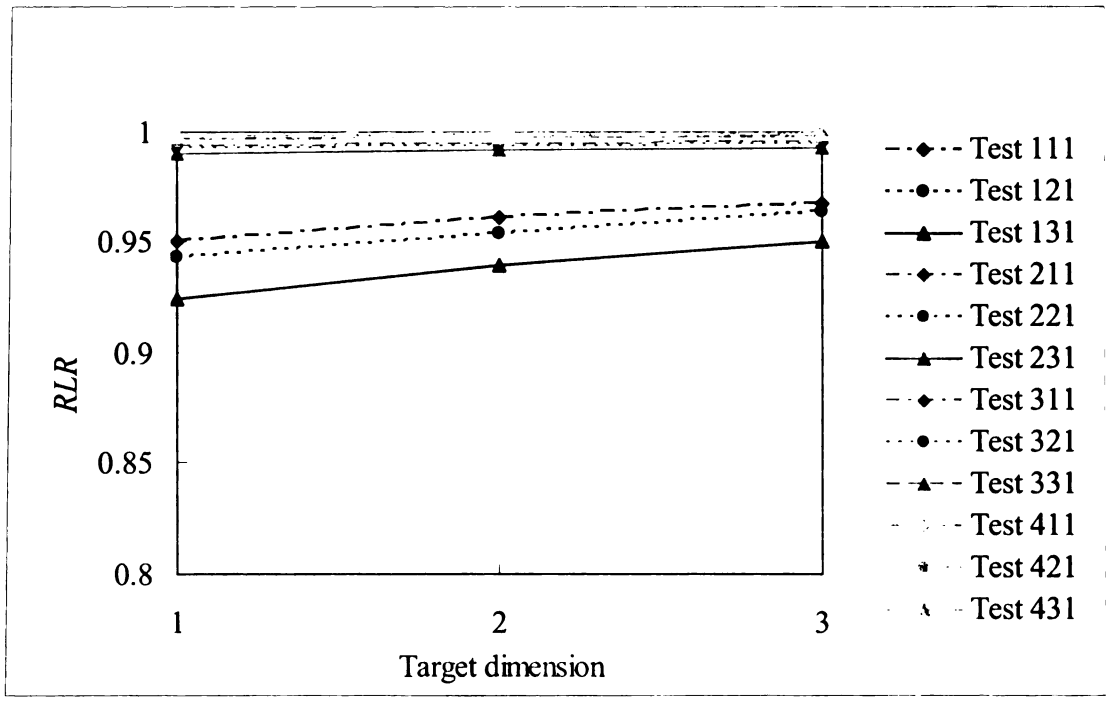


Figure 4.1.2. The change of *RLR* with dimensionality for a 25-item test and 6000 examinees

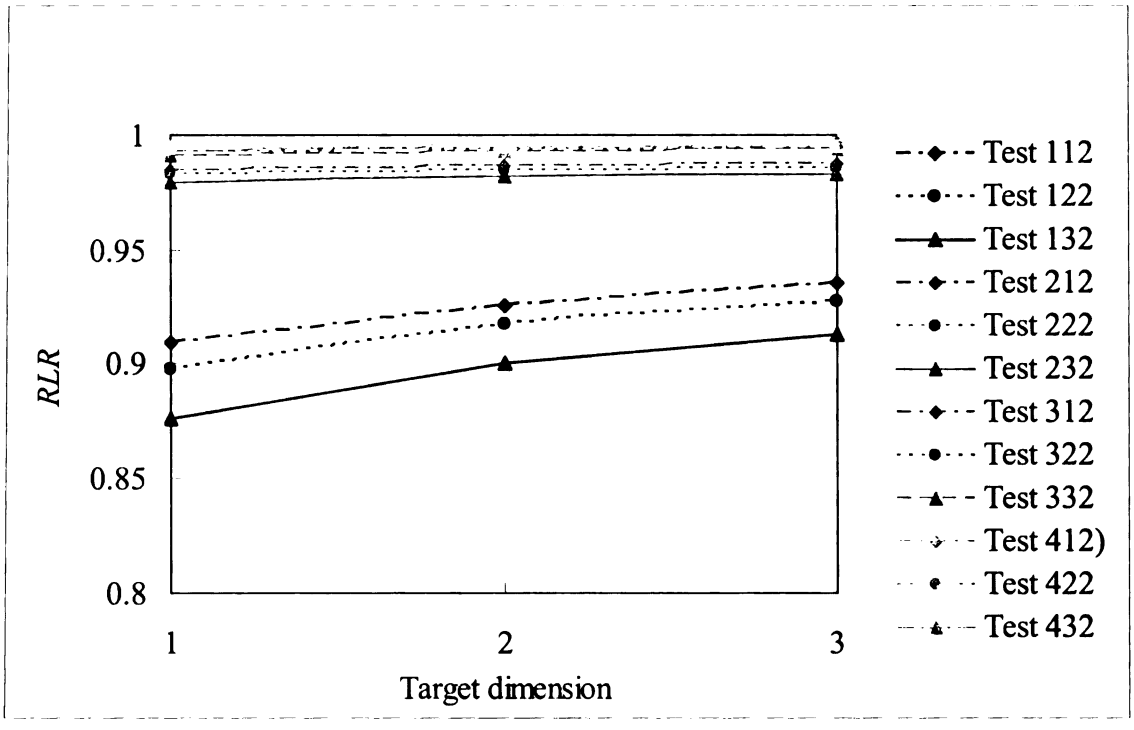


Figure 4.1.3. The change of *RLR* with dimensionality for a 50-item test and 2000 examinees

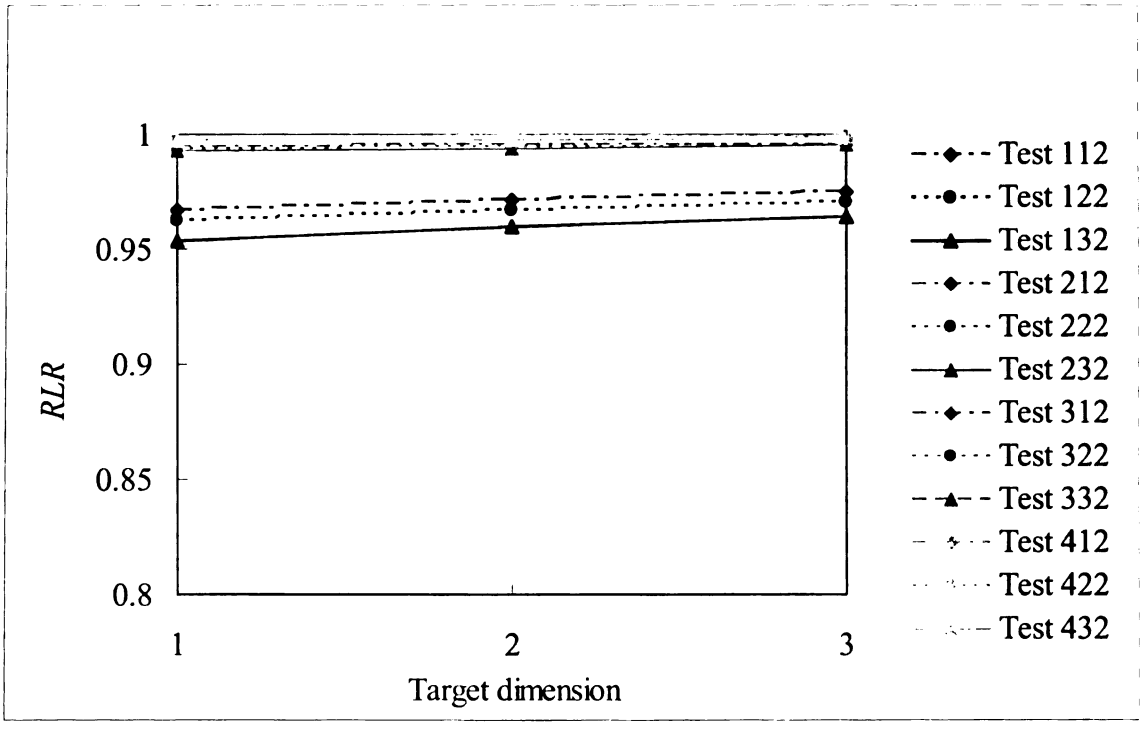


Figure 4.1.4. The change of *RLR* with dimensionality for a 50-item test and 6000 examinees

4.1.2 Results of Multivariate Analysis of Variance for Study I

To explore the influence of manipulated factors on the *RLR* index, a Multivariate Analysis of Variance (MANOVA) was conducted. The dependent variables in the MANOVA model were the *RLR* indices representing three levels of dimensionality (RLR_1, RLR_2, and RLR_3), and the independent variables were *A*, *D*, *S*, and *T*.

To test whether the overall multivariate difference was significant , Pillai's Trace was employed because it is more robust than other statistics (Wilks' λ, Hotelling's T^2, and Roy's greatest characteristic root) when assumptions are not met (Olson, 1976). As Table 4.1.4 shows, the main effects of *A*, *D*, *S*, *T*, and the interactions were all significant so the hypothesis that there was no between-group difference was rejected. Several of these significant factors had substantive effect sizes, such as *A* ($F(9, 13950)= 757.18$, $p < .01$, $\eta^2= .328$), *D* ($F(6, 9298)= 274.68$, $p< .01$, $\eta^2= .151$), *S* ($F(3, 4648)= 6230.61$, $p < .01$, $\eta^2= .801$), *T* ($F(3, 4648)= 580.61$, $p < .01$, $\eta^2= .273$), $A \times D$ ($F(18, 13950)= 124.11$, $p < .0$, $\eta^2= .138$), $A \times S$ ($F(9, 13950)= 613.79$, $p < .01$, $\eta^2= .284$), and $A \times T$ ($F(9, 13950)= 284.63$, $p < .01$, $\eta^2= .155$). They should be considered as having important effects on the *RLR* indices. The interactions $D \times S$, $D \times T$, $S \times T$, $A \times D \times S$, $A \times D \times T$, $A \times S \times T$, $D \times S \times T$, and $A \times D \times S \times T$ were significant, but their effect sizes were small. Because the total number of simulated data sets was 4800, the significance of the interaction terms with small effect sizes may be due to the large sample size in MANOVA. Even though these interactions were significant, they might not have important influence on the dependent valuables.

Table 4.1.4. The multivariate test for Study I

Effect	Value	F	Hypothesis df	Error df	η^2
A	.985	757.18*	9	13950	.328
D	.301	274.68*	6	9298	.151
S	.801	6230.61*	3	4648	.801
T	.273	580.61*	3	4648	.273
$A \times D$	.414	124.11*	18	13950	.138
$A \times S$	.851	613.79*	9	13950	.284
$A \times T$	.465	284.63*	9	13950	.155
$D \times S$	.056	44.99*	6	9298	.028
$D \times T$	.027	21.04*	6	9298	.013
$S \times T$	.054	89.03*	3	4648	.054
$A \times D \times S$	.081	21.40*	18	13950	.027
$A \times D \times T$	.046	11.95*	18	13950	.015
$A \times S \times T$	.106	56.82*	9	13950	.035
$D \times S \times T$	.004	3.44*	6	9298	.002
$A \times D \times S \times T$	.010	2.53*	18	13950	.003

* $p < .01$

Given that the overall difference was significant, the univariate tests for each dependent variable were conducted. First, Levene's test of equality of error variances were all significant (RLR_1: $F(47, 4751) = 128.803$, $p < .01$; RLR_2: $F(47, 4737) = 133.710$, $p < .01$; RLR_3: $F(47, 4650) = 129.233$, $p < .01$), indicating that the variances in different design groups were not homogeneous for each separate ANOVA test. However, Lindman (1974, p. 33) and Box (1954) reported that F statistic is quite robust against the violation of the homogeneity assumption. Since the assumption of equal variance was violated at the .01 level, special caution should be taken when interpreting the results of these separate ANOVA analyses.

Table 4.1.5 summarizes ANOVA tests for RLR_1, RLR_2, and RLR_3. The effect sizes of A, D, S, T, and the interactions were similar for RLR_1, RLR_2, and RLR_3. Again, A, D, S, T, and the interactions $A \times D$, $A \times S$, and $A \times T$ can be considered as having important effects on RLR_1, RLR_2, and RLR_3. A has the largest effect on all the RLR

indices. Moreover, D, S, and T had a smaller effect size than its two-way interaction with A. In RLR_1, for example, D (η^2= .199) < $A \times D$ (η^2= .317); S (η^2= .695) < $A \times S$ (η^2= .781); T (η^2= .250) < $A \times T$ (η^2= .442). These patterns indicated that A was the main variable influencing the RLR indices. To further explore the nature of the interactions, the simple effects were shown in Figure 4.1.5 to Figure 4.1.10.

Table 4.1.5. The univariate test for Study I

Source	*df*	RLR_1			RLR_2			RLR_3		
		MS	*F*	η^2	*MS*	*F*	η^2	*MS*	*F*	η^2
A		2.023	27764.26*	.947	1.194	33861.11*	.956	.833	36971.23*	.960
D	2	.042	579.11*	.199	.022	629.51*	.213	.017	739.77*	.241
S	1	.773	10610.07*	.695	.420	11912.50*	.719	.313	13893.09*	.749
T	1	.113	1547.67*	.250	.036	1019.57*	.180	.013	575.13*	.110
$A \times D$	6	.026	359.70*	.317	.012	339.41*	.305	.008	368.70*	.322
$A \times S$	3	.402	5512.18*	.781	.193	5464.55*	.779	.130	5792.86*	.789
$A \times T$	3	.089	1226.21*	.442	.026	739.64*	.323	.010	433.68*	.219
$D \times S$	2	.008	109.81*	.045	.002	69.44*	.029	.002	81.34*	.034
$D \times T$	2	.004	53.32*	.022	.001	25.14*	.011	.001	31.85*	.014
$S \times T$	1	.019	265.19*	.054	.003	90.14*	.019	.001	43.42*	.009
$A \times D \times S$	6	.004	61.55*	.074	.001	26.73*	.033	.001	27.88*	.035
$A \times D \times T$	6	.002	32.82*	.041	.000	13.55*	.017	.000	13.74*	.017
$A \times S \times T$	3	.013	182.13*	.105	.002	53.85*	.034	.001	25.57*	.016
$D \times S \times T$	2	.001	7.98*	.003	.000	.09	.000	.000	1.35	.001
$A \times D \times S \times T$	6	.000	5.00*	.006	.000	.05	.000	.000	.26	.000
Error	4652	.000			.000			.000		
Total	4699									

* p< .01

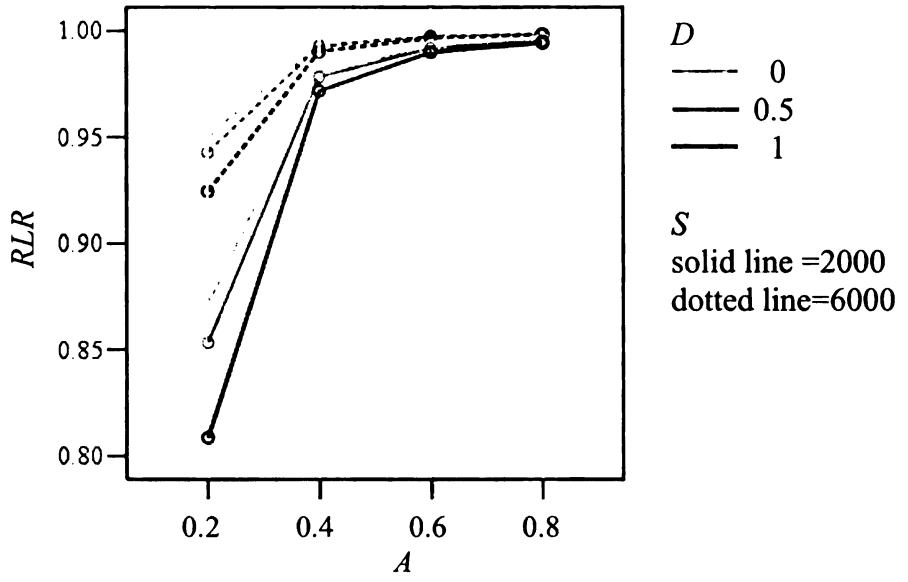


Figure 4.1.5. The interaction of A, D, and S in RLR_1 for 25-item test

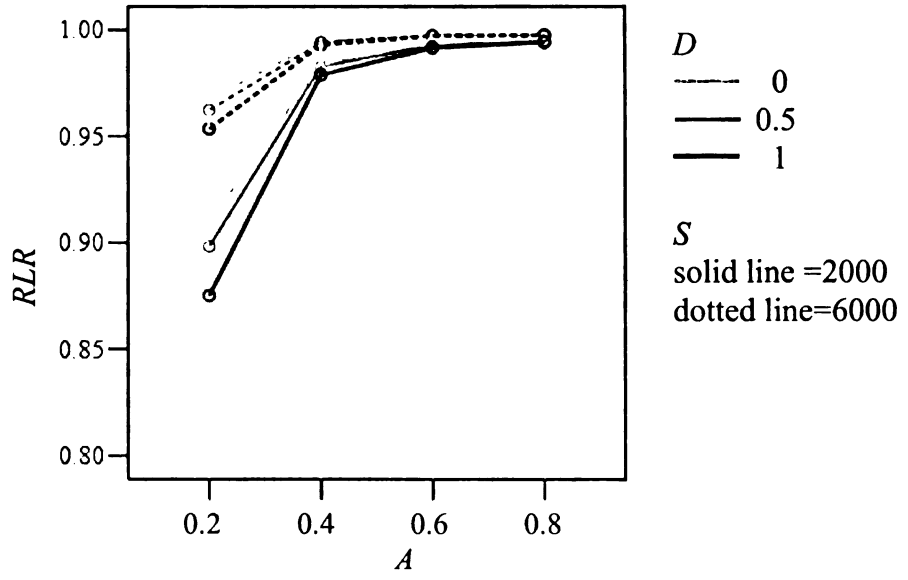


Figure 4.1.6. The interaction of A, D, and S in RLR_1 for 50-item test

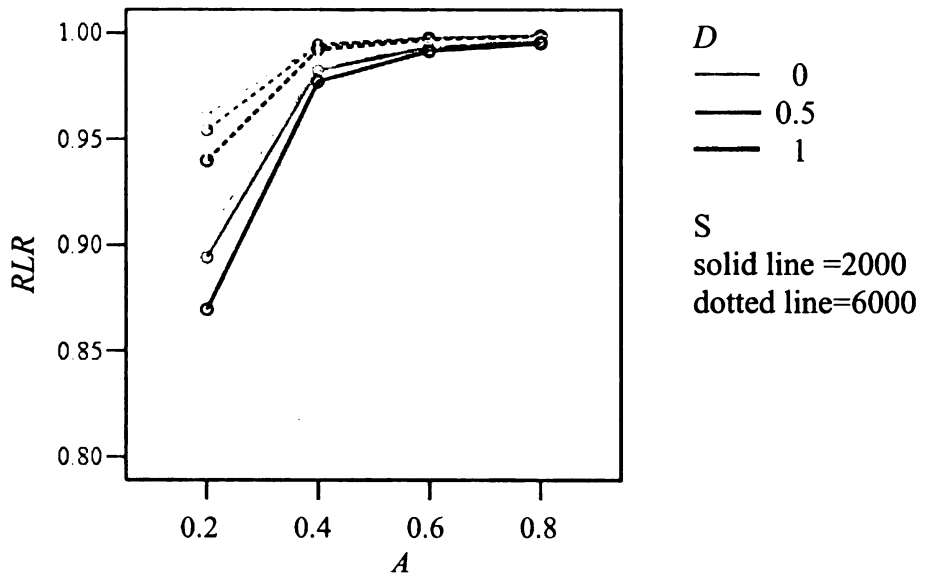


Figure 4.1.7. The interaction of A, D, and S in RLR_2 for 25-item test

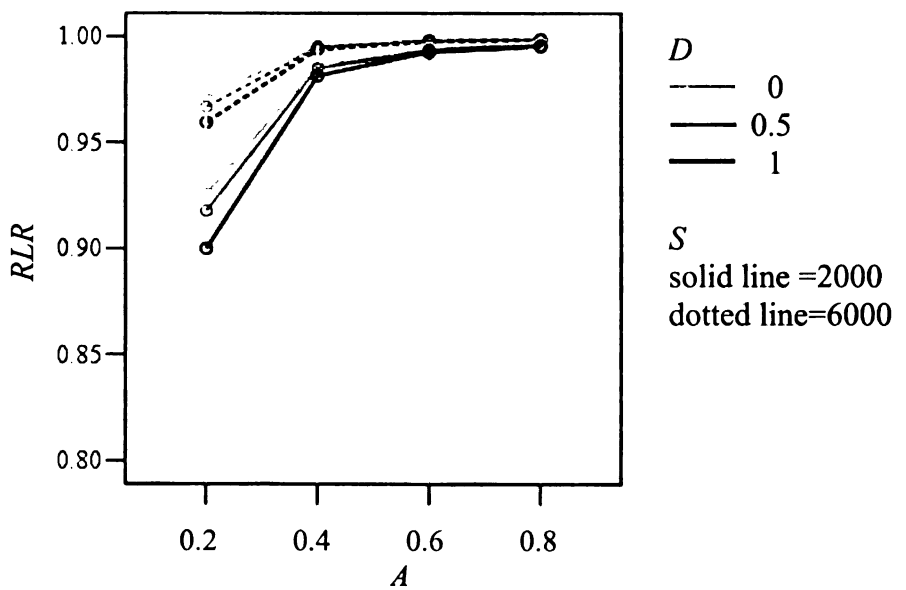


Figure 4.1.8. The interaction of A, D, and S in RLR_2 for 50-item test

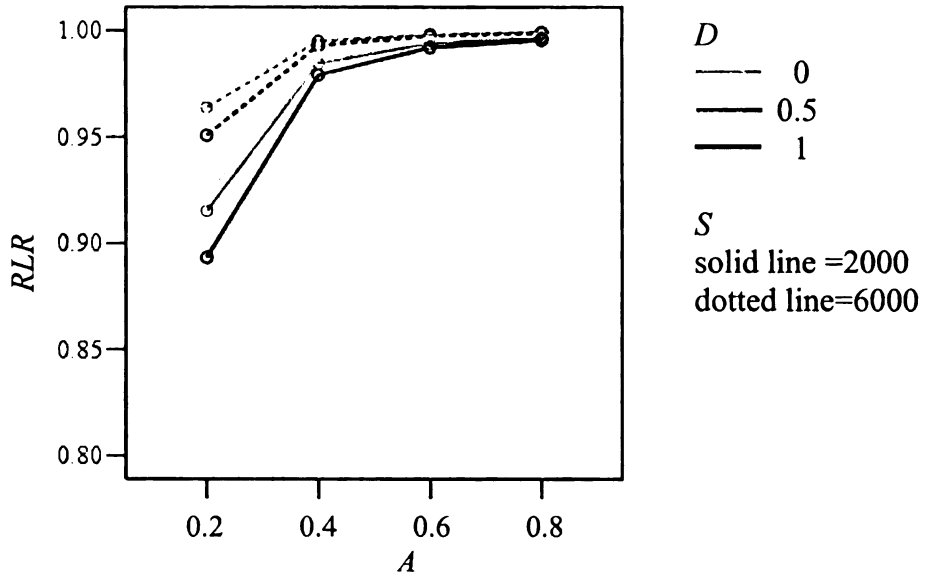


Figure 4.1.9. The interaction of A, D, and S in RLR_3 for 25-item test

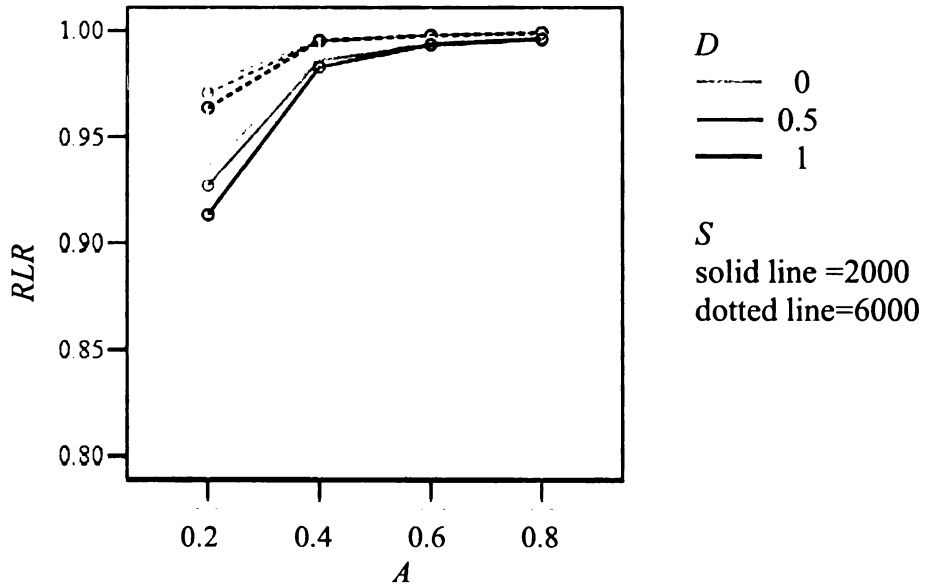


Figure 4.1.10. The interaction of A, D, and S in RLR_3 for 50-item test

Conditioned on T, the patterns of the interactions of A, D, and S were similar across RLR_1, RLR_2, and RLR_3. When the model fit the data, as shown in Figure 4.1.5 and Figure 4.1.6, D had a noticeably negative effect on RLR_1 when A= 0.2. However, the effect of D varied depending on S: when S was 6000, the decrease of RLR_1 due to the increase of D was small; when S was 2000, the descent of RLR_1 due to the increase of D was great. When A= 0.4, the negative effect of D was not obvious, especially when S= 6000. When A= 0.6 or 0.8, the effect of D was minor and thus only the effect of S could be identified.

When the model over-fit the data, as shown in Figure 4.1.7 to Figure 4.1.10, it is clear that D had an effect on RLR_2 and RLR_3, but the effect varied dependeing on A. As long as A was greater than 0.2, the effect of D was minor. Moreover, S still had an effect on RLR_2 and RLR_3, but varied depending on the level of A: the effect of S was great when A was small, but small when A was great.

4.1.3 Comparisons of the Numbers of Rejections

This part of the analysis involved comparing the empirical Type I error rate of the *RLR* index with those from the G^2 test and the G^2_{diff} test in different testing conditions. The theoretical α used for the G^2 test and the G^2_{diff} test was .05. The results in Table 4.1.6 show that the G^2 test at all times rejected the true model regardless of the levels of A, D, S, and T. The results of the G^2_{diff} test were not satisfactory either. With a short T and a small S, the minimum number of rejections was 68 out of 100. Given the same levels of A, D, and T, a large S didn't help decreasing the

number of false rejections. When T was long, the minimum number of rejections was 98 out of 100, regardless of S. A large T tended to inflate the number of rejections more severely than a large S. These results were indicative of a severely inflated Type I error rate problem of using the G^2 test and the G^2_{diff} test to determine whether or not the test data were unidimensional.

The number of rejections of the *RLR* index was computed based on the linear regression technique. With the information of the estimated a-parameters (EA), estimated d-parameters (ED), sample size (S), and test length (T), the lower bound of a good fit for the unidimensional model can be predicted. With the item parameter estimates obtained in Study I, the unidimensional regression model (adjusted R^2 equal to .709) can be expressed as

$$RLR_1 = 0.817509 + 0.000021(S) + 0.001251(T) - 0.020432(ED) + 0.050065(EA) + 0.000023(EA \times S) - 0.001083(EA \times T) + 0.067449(EA \times ED) + 0.000000166(EA \times S \times T). \quad (41)$$

If the observed RLR_1 was smaller than the lower bound, the null hypothesis H_0: d= 1 was rejected. As shown in Table 4.1.6, when S= 2000, the numbers of the false rejections were high for Test 111, Test 121, Test 131, and Test 132, indicating that the low level of A inflated the Type I error rate. Given that A= 0.2 and S= 2000, the number of false rejections inflated with the increase of D. When A= 0.2 and S= 6000, all the false rejections were less than 5 regardless of the levels of D and T. Conversely, for the cases when A was equal to or greater than 0.4, the numbers of rejections were low regardless of the levels of D, S, and T.

Comparing the numbers of false rejections for the three indices under different testing conditions, the *RLR* index outperformed the G^2 test and the G^2_{diff} test. A large sample size and a long test both inflated the Type I error rates for the G^2 test and the G^2_{diff} test, but helped reducing the Type I error rates for the *RLR* index.

Table 4.1.6. The number of rejections in 100 replications for unidimensional data

Data sets	2000 examinees			6000 examinees		
	RLR	G^2	G^2_{diff}	*RLR*	G^2	G^2_{diff}
25-item test						
Test 111	29	100	74	0	100	83
Test 121	33	100	80	0	100	80
Test 131	76	100	82	3	100	67
Test 211	0	100	71	0	100	73
Test 221	0	100	72	0	100	80
Test 231	0	100	75	0	100	74
Test 311	0	100	80	0	100	69
Test 321	0	100	75	0	100	75
Test 331	0	100	68	0	100	67
Test 411	0	100	79	0	100	87
Test 421	0	100	74	0	100	75
Test 431	0	100	75	0	100	68
50-item test						
Test 112	4	100	100	0	100	100
Test 122	2	100	100	0	100	98
Test 132	18	100	100	0	100	100
Test 212	0	100	100	0	100	100
Test 222	0	100	98	0	100	98
Test 232	0	100	100	0	100	100
Test 312	0	100	100	0	100	100
Test 322	0	100	100	0	100	100
Test 332	0	100	100	0	100	97
Test 412	0	100	100	0	100	100
Test 422	0	100	100	0	100	100
Test 432	0	100	100	0	100	100

4.2 Simulation Study II (Multidimensional Data Sets)

The purpose of Study II was to investigate how well the *RLR* index determined the dimensionality for multidimensional data. Again, when the simulated data were analyzed by different levels of multidimensional MIRT models, some of the TESTFACT runs failed because the data generated a singular tetrachoric correlation matrix. Table 4.2.1 shows the number of unsuccessful runs out of 100 replications for each condition. The two-dimensional data had higher rates of unsuccessful TESTFACT runs for the four-dimensional model than for the five-dimensional model, whereas the three-dimensional data had higher rates of unsuccessful TESTFACT runs for the five-dimensional model than for the four-dimensional model.

Given the same levels of *C* and *I*, the rate of getting a singular tetrachoric correlation matrix was high when *A* was moderate. Conditioned on *A* and *C*, the third level of *I* (36: 6: 6) generated a singular tetrachoric correlation matrix at lower rates than the first level (12: 12: 24) and second level (16: 16: 16) of *I*.

Table 4.2.1. The number of unsuccessful TESTFACT runs in Study II

correlation matrix	Form	MIRT model				
		1 Factor	2 Factor	3 Factor	4 Factor	5 Factor
C_1						
	Form 111	0	0	0	32	4
	Form 112	0	0	0	0	2
	Form 121	0	0	0	21	8
	Form 122	0	0	0	1	0
	Form 131	0	0	0	3	1
	Form 132	0	0	0	0	0
C_2						
	Form 211	0	0	0	29	5
	Form 212	0	0	0	0	1
	Form 221	0	0	0	24	12
	Form 222	0	0	0	2	0
	Form 231	0	0	0	6	2
	Form 232	0	0	0	1	0
C_3						
	Form 311	0	0	0	29	9
	Form 312	0	0	0	2	4
	Form 321	0	0	0	30	16
	Form 322	0	0	0	2	1
	Form 331	0	0	0	11	4
	Form 332	0	0	0	2	3
C_4						
	Form 411	0	0	0	0	17
	Form 412	0	0	0	0	0
	Form 421	0	0	0	0	17
	Form 422	0	0	0	0	3
	Form 431	0	0	0	0	3
	Form 432	0	0	0	0	0
C_5						
	Form 511	0	0	0	0	24
	Form 512	0	0	0	0	0
	Form 521	0	0	0	0	19
	Form 522	0	0	0	0	2
	Form 531	0	0	0	0	3
	Form 532	0	0	0	0	2
C_6						
	Form 611	0	0	0	0	20
	Form 612	0	0	0	0	0
	Form 621	0	0	0	0	17
	Form 622	0	0	0	0	0
	Form 631	0	0	0	0	1
	Form 632	0	0	0	0	0

4.2.1 Results of the Summary Statistics

Table 4.2.2 and Table 4.2.3 tabulate the summary statistics of the *RLR* values for

each combination. In order to show the change of *RLR* values associated with dimensionality, Figure 4.2.1 to Figure 4.2.6 were provided with colors denoting different levels of *I* and the line shapes representing different levels of *A*. Besides, the conditional distributions of the *RLR* values were offered in Appendix C as a supplement to the summary statistics.

In Table 4.2.2 and Table 4.2.3, some of the *RLR* values slightly exceeded 1 when the model recovered the true dimensionality or over-fit the data. The unexpected *RLR* values showed that the lower-factor model fit the data better than the higher-factor model. However, for every case when the *RLR* values exceeded 1, a negative G^2_{diff} statistic occurred. A negative value of the G^2_{diff} statistic indicated that the discrepancy between the predicted frequency and the observed frequency for the lower-factor model is smaller than that of the higher-factor model. The discussion of the occurrence of the unexpected values for the *RLR* index and the G^2_{diff} test was provided in Chapter 5.

Again, the *SD* of the *RLR* values in each condition was small. Conditioned on *A*, *C*, and *I*, the *SD* of the *RLR* values was great when the model under-fit the data. Given the same levels of *C* and *I*, RLR_I was low when *A* was high. With the same levels of *C* and *A*, RLR_I was high when the dominant factor was strong. For the data generated with the correlation matrices C_1, C_2, and C_3, the *RLR* values approached 1 for the two-dimensional model, and did not obviously increase for the higher-factor models. For the data generated with the correlation matrices C_4, C_5, and C_6, the *RLR* values approached 1 for the three-dimensional model, and did not increase for the four-dimensional model. In general, the patterns of the *RLR* values reflected the simulated dimensionality.

Table 4.2.2. Summary statistics of the *RLR* index for two-dimensional data sets

Form	*RLR*	Descriptive statistics				Test	Descriptive statistics			
		Mean	*SD*	*N*	*SE*		Mean	*SD*	*N*	*SE*
Form 111						Form 112				
	RLR_1	0.8904	0.0073	100	0.0007		0.8380	0.0079	100	0.0008
	RLR_2	0.9951	0.0011	100	0.0001		1.0003	0.0009	100	0.0001
	RLR_3	0.9954	0.0015	68	0.0002		0.9990	0.0013	100	0.0001
	RLR_4	0.9930	0.0012	66	0.0001		0.9924	0.0012	99	0.0001
Form 121						Form 122				
	RLR_1	0.8954	0.0077	100	0.0008		0.8727	0.0061	100	0.0006
	RLR_2	0.9940	0.0018	100	0.0002		0.9990	0.0008	100	0.0001
	RLR_3	0.9938	0.0019	76	0.0002		0.9984	0.0012	99	0.0001
	RLR_4	0.9926	0.0015	74	0.0002		0.9932	0.0012	99	0.0001
Form 131						Form 132				
	RLR_1	0.9725	0.0032	100	0.0003		0.9547	0.0027	100	0.0003
	RLR_2	0.9933	0.0012	100	0.0001		0.9980	0.0007	100	0.0001
	RLR_3	0.9935	0.0013	97	0.0001		0.9990	0.0010	100	0.0001
	RLR_4	0.9939	0.0010	96	0.0001		0.9950	0.0010	100	0.0001
Form 211						Form 212				
	RLR_1	0.7276	0.0122	100	0.0012		0.6453	0.0119	100	0.0012
	RLR_2	0.9959	0.0015	100	0.0001		1.0015	0.0016	100	0.0002
	RLR_3	0.9955	0.0018	71	0.0002		0.9993	0.0021	100	0.0002
	RLR_4	0.9921	0.0017	67	0.0002		0.9904	0.0013	99	0.0001
Form 221						Form 222				
	RLR_1	0.7305	0.0136	100	0.0014		0.7357	0.0092	100	0.0009
	RLR_2	0.9944	0.0016	100	0.0002		1.0000	0.0014	100	0.0001
	RLR_3	0.9941	0.0018	76	0.0002		0.9989	0.0016	98	0.0002
	RLR_4	0.9914	0.0017	69	0.0002		0.9915	0.0013	98	0.0001
Form 231						Form 232				
	RLR_1	0.9413	0.0046	100	0.0005		0.9170	0.0039	100	0.0004
	RLR_2	0.9930	0.0016	100	0.0002		0.9982	0.0008	100	0.0001
	RLR_3	0.9940	0.0018	94	0.0002		0.9997	0.0010	99	0.0001
	RLR_4	0.9931	0.0011	92	0.0001		0.9938	0.0011	99	0.0001
Form 311						Form 312				
	RLR_1	0.6049	0.0118	100	0.0012		0.5011	0.0131	100	0.0013
	RLR_2	0.9952	0.0013	100	0.0001		1.0012	0.0012	100	0.0001
	RLR_3	0.9963	0.0015	71	0.0002		1.0001	0.0017	98	0.0002
	RLR_4	0.9919	0.0016	65	0.0002		0.9893	0.0015	95	0.0002
Form 321						Form 322				
	RLR_1	0.6025	0.0121	100	0.0012		0.6586	0.0088	100	0.0009
	RLR_2	0.9936	0.0015	100	0.0002		0.9988	0.0012	100	0.0001
	RLR_3	0.9952	0.0017	70	0.0002		0.9999	0.0013	98	0.0001
	RLR_4	0.9910	0.0017	58	0.0002		0.9905	0.0013	97	0.0001
Form 331						Form 332				
	RLR_1	0.9240	0.0055	100	0.0006		0.8988	0.0040	100	0.0004
	RLR_2	0.9930	0.0013	100	0.0001		0.9978	0.0013	100	0.0001
	RLR_3	0.9943	0.0015	89	0.0002		0.9996	0.0012	98	0.0001
	RLR_4	0.9924	0.0012	85	0.0001		0.9932	0.0012	95	0.0001

Table 4.2.3. Summary statistics of the *RLR* index for three-dimensional data sets

Form	*RLR*	Descriptive statistics				Form	Descriptive statistics			
		Mean	*SD*	*N*	*SE*		Mean	*SD*	*N*	*SE*
Form 411						Form 412				
	RLR_1	0.8382	0.0101	100	0.0010		0.8050	0.0094	100	0.0009
	RLR_2	0.9574	0.0050	100	0.0005		0.9144	0.0060	100	0.0006
	RLR_3	0.9961	0.0016	100	0.0002		0.9990	0.0018	100	0.0002
	RLR_4	0.9904	0.0017	83	0.0002		0.9853	0.0020	100	0.0002
Form 421						Form 422				
	RLR_1	0.8177	0.0124	100	0.0012		0.7542	0.0116	100	0.0012
	RLR_2	0.9092	0.0081	100	0.0008		0.8864	0.0082	100	0.0008
	RLR_3	0.9959	0.0025	100	0.0003		0.9994	0.0021	100	0.0002
	RLR_4	0.9888	0.0018	83	0.0002		0.9848	0.0021	97	0.0002
Form 431						Form 432				
	RLR_1	0.9558	0.0050	100	0.0005		0.9252	0.0045	100	0.0005
	RLR_2	0.9676	0.0042	100	0.0004		0.9512	0.0037	100	0.0004
	RLR_3	0.9932	0.0015	100	0.0002		0.9987	0.0014	100	0.0001
	RLR_4	0.9919	0.0014	97	0.0001		0.9900	0.0014	100	0.0001
Form 511						Form 512				
	RLR_1	0.7540	0.0124	100	0.0012		0.6690	0.0123	100	0.0012
	RLR_2	0.9435	0.0062	100	0.0006		0.9106	0.0065	100	0.0006
	RLR_3	0.9962	0.0022	100	0.0002		0.9978	0.0025	100	0.0002
	RLR_4	0.9883	0.0019	76	0.0002		0.9818	0.0019	100	0.0002
Form 521						Form 522				
	RLR_1	0.6366	0.0186	100	0.0019		0.6318	0.0151	100	0.0015
	RLR_2	0.8902	0.0081	100	0.0008		0.8456	0.0086	100	0.0009
	RLR_3	0.9956	0.0032	100	0.0003		0.9981	0.0023	100	0.0002
	RLR_4	0.9870	0.0022	81	0.0002		0.9823	0.0019	98	0.0002
Form 531						Form 532				
	RLR_1	0.9138	0.0065	100	0.0007		0.8838	0.0058	100	0.0006
	RLR_2	0.9698	0.0041	100	0.0004		0.9505	0.0045	100	0.0005
	RLR_3	0.9932	0.0018	100	0.0002		0.9985	0.0023	100	0.0002
	RLR_4	0.9913	0.0015	97	0.0002		0.9876	0.0019	98	0.0002
Form 611						Form 612				
	RLR_1	0.7681	0.0102	100	0.0010		0.7041	0.0098	100	0.0010
	RLR_2	0.8838	0.0077	100	0.0008		0.8144	0.0071	100	0.0007
	RLR_3	0.9957	0.0034	100	0.0003		0.9946	0.0043	100	0.0004
	RLR_4	0.9847	0.0023	80	0.0003		0.9778	0.0019	100	0.0002
Form 621						Form 622				
	RLR_1	0.5915	0.0154	100	0.0015		0.4847	0.0218	100	0.0022
	RLR_2	0.7398	0.0120	100	0.0012		0.6688	0.0222	100	0.0022
	RLR_3	0.9934	0.0051	100	0.0005		0.9929	0.0036	100	0.0004
	RLR_4	0.9845	0.0025	83	0.0003		0.9783	0.0031	100	0.0003
Form 631						Form 632				
	RLR_1	0.9126	0.0088	100	0.0009		0.8865	0.0065	100	0.0007
	RLR_2	0.9450	0.0082	100	0.0008		0.9121	0.0060	100	0.0006
	RLR_3	0.9948	0.0023	99	0.0002		1.0002	0.0025	100	0.0002
	RLR_4	0.9887	0.0023	99	0.0002		0.9838	0.0018	100	0.0002

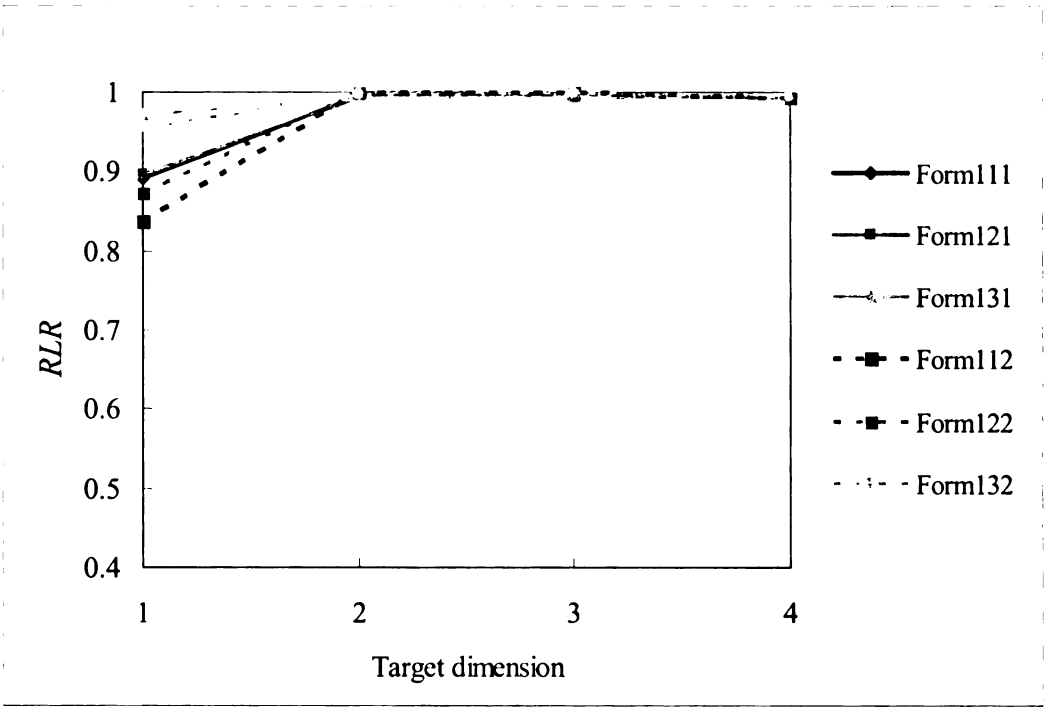


Figure 4.2.1. The change of *RLR* with dimensionality for the correlation matrix C_1

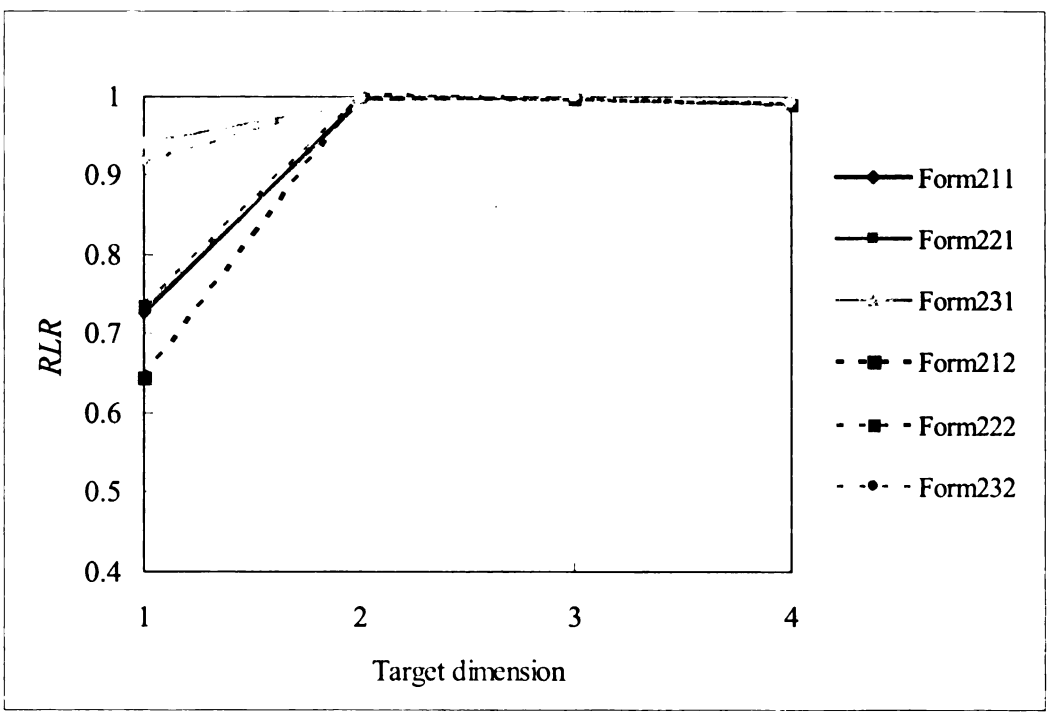


Figure 4.2.2. The change of *RLR* with dimensionality for the correlation matrix C_2

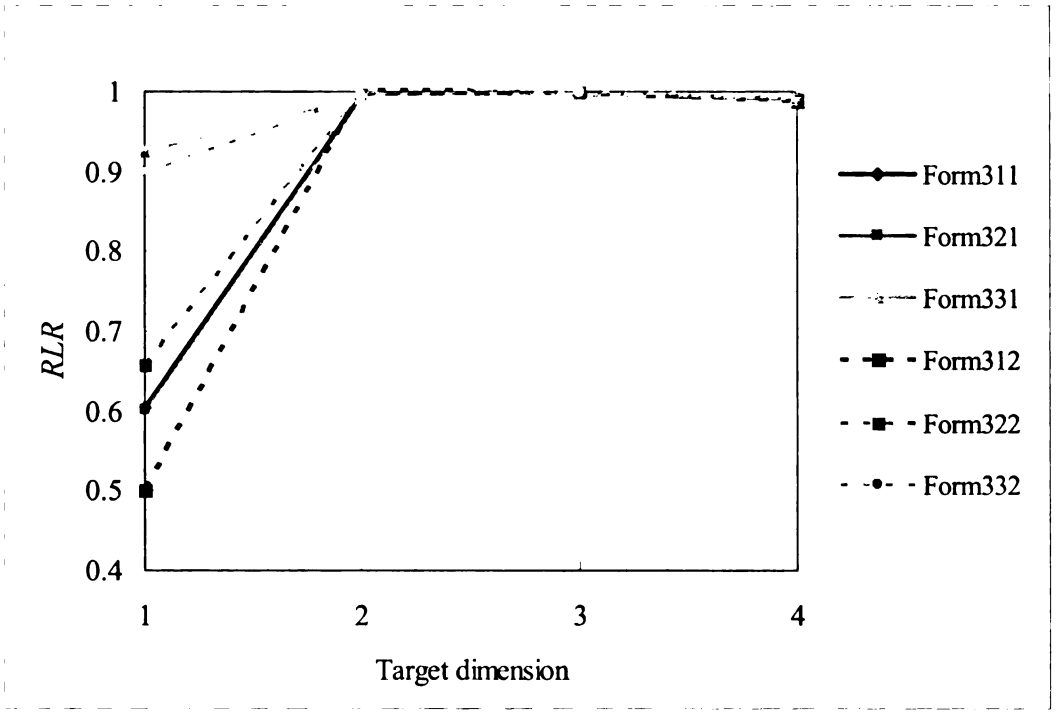


Figure 4.2.3. The change of *RLR* with dimensionality for the correlation matrix C_3

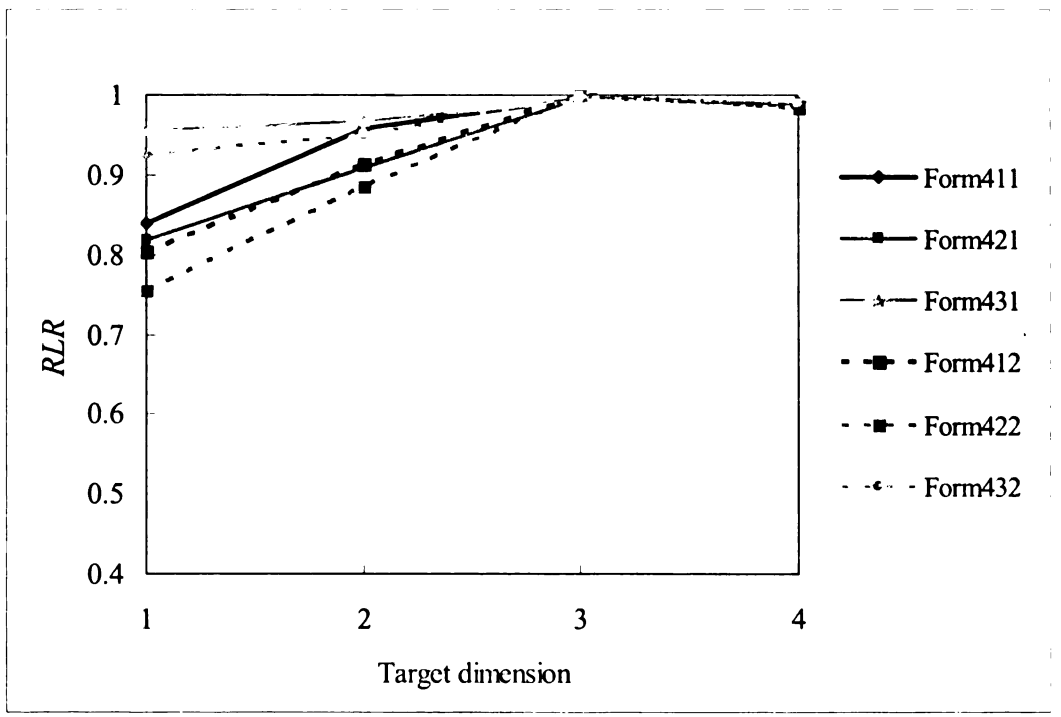


Figure 4.2.4. The change of *RLR* with dimensionality for the correlation matrix C_4

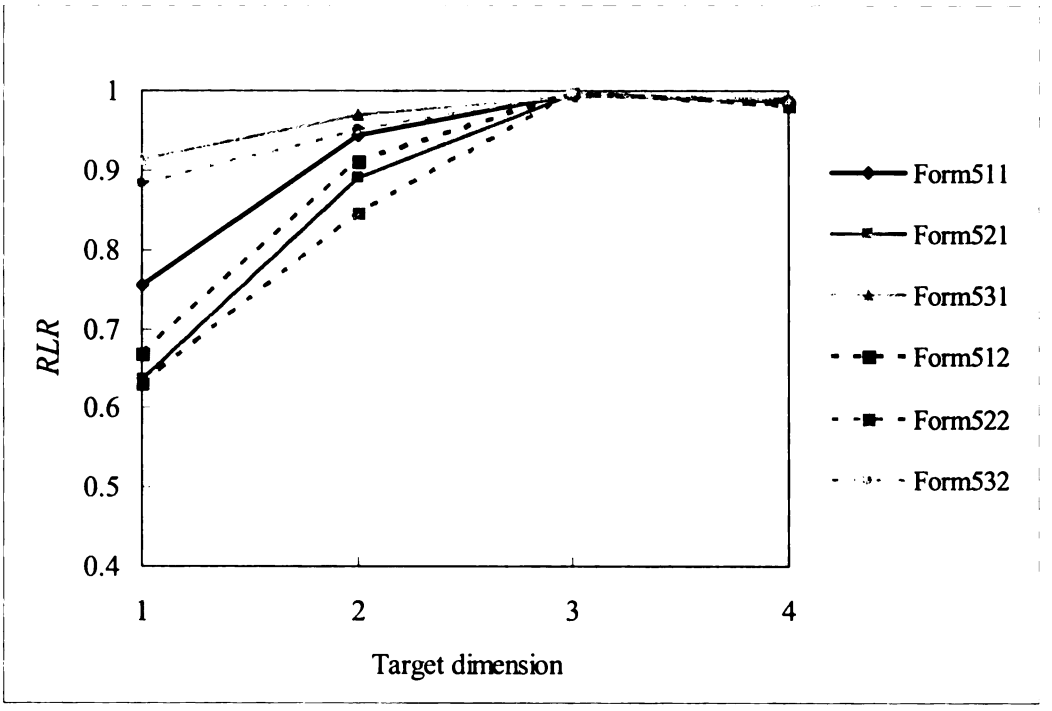


Figure 4.2.5. The change of *RLR* with dimensionality for the correlation matrix C_5

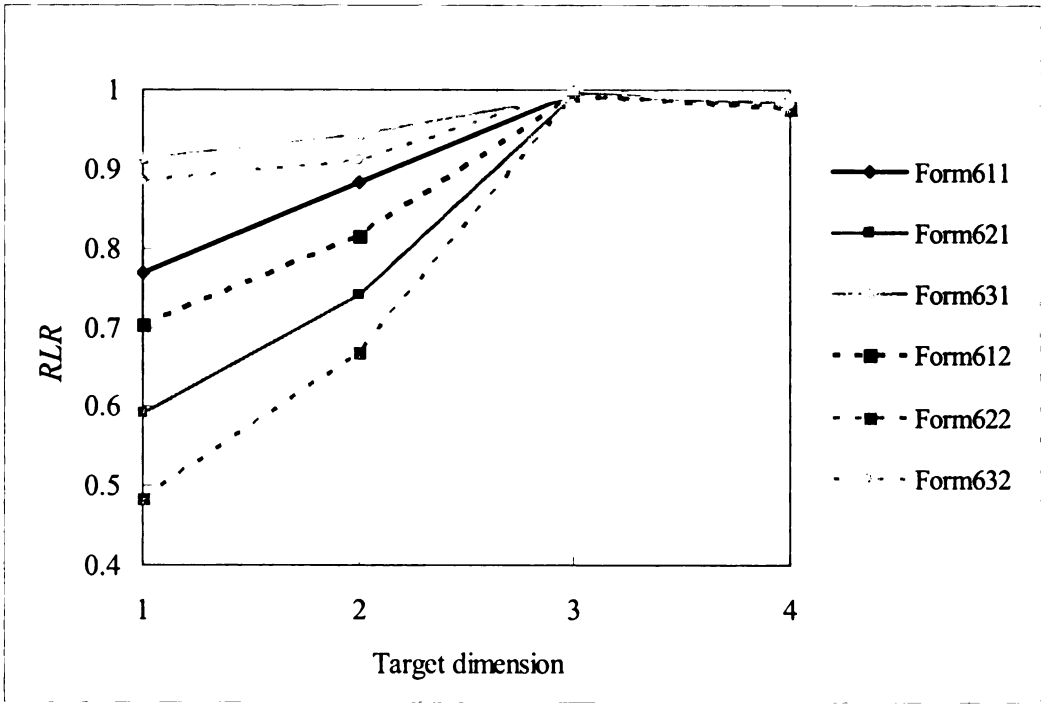


Figure 4.2.6. The change of *RLR* with dimensionality for the correlation matrix C_6

4.2.2 Results of Multivariate Analysis of Variance for Study II

Again, a MANOVA analysis was conducted to explore the influence of the manipulated factors on the *RLR* index. The dependent variables in the MANOVA model were the *RLR* indices representing four levels of dimensionality (RLR_1, RLR_2, RLR_3, and RLR_4), and the independent variables were *A*, *C*, and *I*. Again, the Pillai's Trace was employed to test the overall multivariate difference because of its robustness to the violation of the assumption of homogeneity of variance.

Table 4.2.4. The multivariate test for Study II

Effect	Value	*F*	Hypothesis *df*	Error *df*	η^2
A	0.870	5310.849*	4	3182	0.870
C	2.079	689.733*	20	12740	0.520
I	1.799	7116.803*	8	6366	0.899
A×*C*	0.913	188.489*	20	12740	0.228
A×*I*	0.740	466.939*	8	6366	0.370
C×*I*	2.021	325.152*	40	12740	0.505
A×*C*×*I*	0.989	104.627*	40	12740	0.247

* $p < .01$

The overall multivariate test shown in Table 4.2.4 was significant, indicating that there was a significant difference overall for the main effects *A*, *C*, *I*, and the interactions on the *RLR* indices representing different levels of dimensionality. Based on the results of the significance test and effect size, *A*, *C*, *I* and *C*×*I* had important effects on the *RLR* indices: *A* ($F(4, 3182) = 5310.849$, $p < .01$, $\eta^2 = 0.870$), *C* ($F(20, 12740) = 689.733$, $p < .01$, $\eta^2 = 0.520$), *I* ($F(8, 6366) = 7116.803$, $p < .01$, $\eta^2 = 0.899$), and *C*×*I* ($F(40, 12740) = 325.152$, $p < .01$, $\eta^2 = 0.505$). The remaining interactions had relatively minor effects on the *RLR* indices: *A*×*C* ($F(20, 12740) = 188.489$, $p < .01$, $\eta^2 = 0.228$), *A*×*I* ($F(8, 6366) = 466.939$,

$p< .01, \eta^2= 0.370$), and $A \times C \times I$ ($F(40, 12740)= 104.627$, $p< .01$, $\eta^2= 0.247$).

Because the overall difference was significant, the effects of A, C, and I on RLR_1, RLR_2, RLR_3, and RLR_4 can be explored by separate univariate analysis. First, Levene's test of equality of error variances were all significant (RLR_1: $F(35, 3564)= 31.923$, $p< .01$; RLR_2: $F(35, 3564)= 131.726$, $p< .01$; RLR_3: $F(35, 3365)= 12.661$, $p< .01$); RLR_4: $F(35, 3189)= 8.820$, $p< .01$). Since the assumption of homoscedasticity for the four separate univariate tests were all violated at the .01 level, attention should be paid when interpreting the univariate analyses.

Table 4.2.5. The univariate test for Study II

Source	*df*	RLR_1			RLR_2		
		MS	*F*	η^2	*MS*	*F*	η^2
A	1	1.405	12904.228*	0.784	0.259	6834.570*	0.657
C	5	3.699	33973.211*	0.979	2.736	72327.056*	0.990
I	2	17.690	162488.505*	0.989	1.210	31981.079*	0.947
A×*C*	5	0.031	282.963*	0.284	0.105	2767.152*	0.795
A×*I*	2	0.211	1941.125*	0.521	0.014	378.345*	0.175
C×*I*	10	0.970	8913.046*	0.962	0.401	10602.992*	0.967
A×*C*×*I*	10	0.081	747.823*	0.677	0.005	140.905*	0.283
Error	3564	0.000			0.000		
Total	3600						

(Continued)

Source	*df*	RLR_3			RLR_4		
		MS	*F*	η^2	*MS*	*F*	η^2
A	1	0.012	2279.148*	0.404	0.005	1755.672*	0.355
C	5	0.000	68.075*	0.092	0.008	2827.289*	0.816
I	2	0.000	33.958*	0.020	0.004	1270.387*	0.443
A×*C*	5	0.000	51.470*	0.071	0.001	320.215*	0.334
A×*I*	2	0.001	133.720*	0.074	0.000	150.458*	0.086
C×*I*	10	0.000	47.435*	0.124	0.000	46.238*	0.127
A×*C*×*I*	10	0.000	11.962*	0.034	0.000	2.926*	0.009
Error	3564	0.000			0.000		
Total	3600						

* $p< .01$

Based on the results shown in Table 4.2.5, all the main effects and interactions were significant, but their effects on RLR_1, RLR_2, RLR_3 and RLR_4 were different. The effect size of A decreased from RLR_1 to RLR_4. However, C and I had large effect sizes for RLR_1, RLR_2, but a small effect size for RLR_4 and the smallest effect size for RLR_3. Concerning the interaction $A \times C$, it had a large effect size for RLR_2 (η^2= .795), moderate effect sizes for RLR_1 (η^2= .284) and RLR_4 (η^2= .334), but a small effect size for RLR_3 (η^2= .071). The interaction $A \times I$ had a moderate effect size for RLR_1 (η^2= .521), and small effect sizes for RLR_2 (η^2= .174), RLR_3 (η^2= .074), and RLR_4 (η^2= .086). The interaction $C \times I$ showed a different pattern. The effect sizes were large for RLR_1 (η^2= .962) and RLR_2 (η^2= .967), but small for RLR_3 (η^2= .124) and RLR_4 (η^2= .127).

These unsystematic changes in the effect sizes for the *RL*R indices were hard to explain. In order to clarify the effect of the manipulated factors on dimensionality, the overall data set was separated into two-dimensional data and three-dimensional data, and again analyzed by MANOVA, respectively. Table 4.2.6 and Table 4.2.7 display multivariate test results based on Pillai's Trace for the two- and three-dimensional data, respectively. For the two-dimensional data, A was the most important variable and had an effect size of .941. The effect sizes of C (η^2= .517), I (η^2= .628) and $A \times I$ (η^2= .492) were moderate, but the effect size of $A \times C$ (η^2= .044) was small. With respect to the three-dimensional data, the effect sizes of A (η^2= .922), C (η^2= .918), I (η^2= .825) were all large. All the interactions were significant with moderate effect sizes.

Table 4.2.6. The multivariate test for the two-dimensional data

Effect	Value	F	Hypothesis df	Error df	η^2
A	0.941	6141.303*	4	1529	0.941
C	1.034	409.833*	8	3060	0.517
I	1.255	644.431*	8	3060	0.628
$A \times C$	0.087	17.432*	8	3060	0.044
$A \times I$	0.984	370.166*	8	3060	0.492
$C \times I$	0.993	126.528*	16	6128	0.248
$A \times C \times I$	0.626	71.044*	16	6128	0.156

* $p < .01$

Table 4.2.7. The multivariate test for the three-dimensional data

Effect	Value	F	Hypothesis df	Error df	η^2
A	0.922	4904.996*	4	1650	0.922
C	1.835	4594.182*	8	3302	0.918
I	1.650	1946.650*	8	3302	0.825
$A \times C$	0.535	150.647*	8	3302	0.267
$A \times I$	0.551	157.017*	8	3302	0.276
$C \times I$	1.502	248.356*	16	6612	0.375
$A \times C \times I$	0.676	84.021*	16	6612	0.169

* $p < .01$

To further determine the nature of the effect, the univariate tests for the two- and three-dimensional data were conducted. Levene's tests of equality of error variances were all significant at the .01 level (RLR_1: $F(17, 1782) = 28.292$, $p < .01$; RLR_2: $F(17, 1782) = 6.326$, $p < .01$; RLR_3: $F(17, 1584) = 4.725$, $p < .01$; RLR_4: $F(17, 1535) = 4.031$, $p < .01$). Levene's tests of equality of error variances for the three-dimensional data were also significant at the .01 level (RLR_1: $F(17, 1782) = 29.881$, $p < .01$; RLR_2: $F(17, 1782) = 71.847$, $p < .01$; RLR_3: $F(17, 17810) = 8.528$, $p < .01$; RLR_4: $F(17, 1654) = 4.1085$, $p < .01$). Even though F test is robust to the violation of the homogeneity assumption, care should be taken when interpreting the following univariate analyses.

Table 4.2.8. Univariate test for two-dimensional data

Source	*df*	RLR_1			RLR_2		
		MS	*F*	η^2	*MS*	*F*	η^2
A	1	0.397	5070.723*	0.740	0.012	7412.411*	0.806
C	2	6.411	81942.183*	0.989	0.000	36.976*	0.040
I	2	9.121	116576.748*	0.992	0.001	638.497*	0.417
A×*C*	2	0.004	45.573*	0.049	0.000	6.489*	0.007
A×*I*	2	0.326	4163.514*	0.824	0.000	12.876*	0.014
C×*I*	4	0.929	11875.816*	0.964	0.000	12.609*	0.028
A×*C*×*I*	4	0.055	705.638*	0.613	0.000	1.629	0.004
Error	1782	0.000			0.000		
Total	1800						

(Continued)

Source	*df*	RLR_3			RLR_4		
		MS	*F*	η^2	*MS*	*F*	η^2
A	1	0.008	3641.697*	0.697	0.000	13.375*	0.009
C	2	0.000	61.309*	0.072	0.001	295.010*	0.278
I	2	0.000	57.118*	0.067	0.001	395.795*	0.340
A×*C*	2	0.000	0.604	0.001	0.000	23.371*	0.030
A×*I*	2	0.000	44.930*	0.054	0.000	114.172*	0.129
C×*I*	4	0.000	5.280*	0.013	0.000	3.037	0.008
A×*C*×*I*	4	0.000	0.173	0.000	0.000	5.228*	0.013
Error	1782	0.000			0.000		
Total	1800						

* $p < .01$

In Table 4.2.8, *A* had large effect sizes of for RLR_1 ($\eta^2 = .740$), RLR_2 ($\eta^2 = .806$), RLR_3 (l $\eta^2 = .697$), but a small effect size for RLR_4 ($\eta^2 = .009$). The effect size of *C* was large for RLR_1 ($\eta^2 = .989$), but dropped to 0.04 for RLR_2 and 0.072 for RLR_3, respectively. The effect size of *I* was large for RLR_1 ($\eta^2 = .992$), but reduced to 0.417 for RLR_2, and then became the smallest for RLR_3 ($\eta^2 = .067$). With regard to Table 4.2.9, all the effect sizes for *A*, *C*, and *I* were small for RLR_3, but large for RLR_1, RLR_2, and RLR_4.

Table 4.2.9. Univariate test for three-dimensional data

Source	*df*	RLR_1			RLR_2		
		MS	*F*	η^2	*MS*	*F*	η^2
A	1	1.095	7848.639*	0.815	0.690	9325.502*	0.840
C	2	2.636	18900.019*	0.955	1.921	25962.423*	0.967
I	2	10.294	73795.933*	0.988	2.444	33037.717*	0.974
A×*C*	2	0.030	215.664*	0.195	0.040	537.825*	0.376
A×*I*	2	0.047	338.056*	0.275	0.030	406.642*	0.313
C×*I*	4	0.634	4544.845*	0.911	0.385	5204.494*	0.921
A×*C*×*I*	10	0.068	484.121*	0.521	0.005	73.352*	0.141
Error	1782	0.000			0.000		
Total	1800						

(Continued)

Source	*df*	RLR_3			RLR_4		
		MS	*F*	η^2	*MS*	*F*	η^2
A	1	0.003	463.401*	0.206	0.010	2418.868*	0.594
C	2	0.001	68.221*	0.071	0.004	1069.574*	0.564
I	2	0.000	10.756*	0.012	0.004	920.176*	0.527
A×*C*	2	0.000	38.279*	0.041	0.000	47.361*	0.054
A×*I*	2	0.001	102.143*	0.103	0.000	62.553*	0.070
C×*I*	4	0.001	75.448*	0.145	0.000	10.951*	0.026
A×*C*×*I*	4	0.000	9.229*	0.020	0.000	2.029	0.005
Error	1782	0.000			0.000		
Total	1800						

* $p < .01$

The different findings for the two- and three-dimensional data reflected the fact that all the data were simulated with a three-dimensional correlation matrice and three-dimensional item parameters. Regarding the two-dimensional data, both the two-dimensional and three-dimensional models should result in a good fit. Thus, the effects of *C*, *I*, *A*×*C*, *A*×*I*, *C*×*I*, and *A*×*I*×*C* were low for RLR_2 and RLR_3. When the model under-fit the two-dimensional data, *A*, *C*, *I* and the interactions were important factors to RLR_1. When the model over-fit the data, only *C*, *I*, and *A*×*I* seemed to affect the size of RLR_4.

For the three-dimensional data, the consistent pattern showed that all the RLR_3

values approached 1 when the model fit the data well. Since the model-data-fit was good, the effects of A, C, and I on the fit index became minor. Conversely, A, C, I and the interactions were all important when the model under-fit the data. When the model over-fit the data, only A, C and I influenced the size of RLR_4.

In order to present the interactions among A, C, and I, the simple effects were displayed in Figure 4.2.7 to Figure 4.2.30. When the model under-fit the data, the RLR value varied depending upon the size of the dominant factor which had the highest percentage of items sensitive to it. In Figure 4.2.7 to Figure 4.2.9, the first level of I (12:12:24) generated the lowest RLR_1 value for the data generated with correlation matrices C_1, C_2, and C_3 because the dominant factor only contained 50% of the items. Conversely, as shown from Figure 4.2.10 to Figure 4.2.12, the second level of I (16:16:16) generated the lowest RLR_1 value for the data generated with correlation matrices C_4, C_5, and C_6, because the dominant factor only contained 33% of the items. Given the same level of C, the distinctions among different levels of I increased when A was high. However, the influence of A was not the same for different combinations of C and I.

Different results about the interactions can be found in Figure 4.2.13 to Figure 4.2.18. For the data generated with correlation matrices C_1, C_2, and C_3, RLR_2 approached 1.00 and implied a good fit. Thus for correlation matrices C_1, C_2, and C_3, the effects of A and I on RLR_2 were minor. With respect to the data generated with correlation matrices C_4, C_5, and C_6, RLR_2 still varied depending on the levels of A, C and I. The effects of A and I on RLR_2 were important.

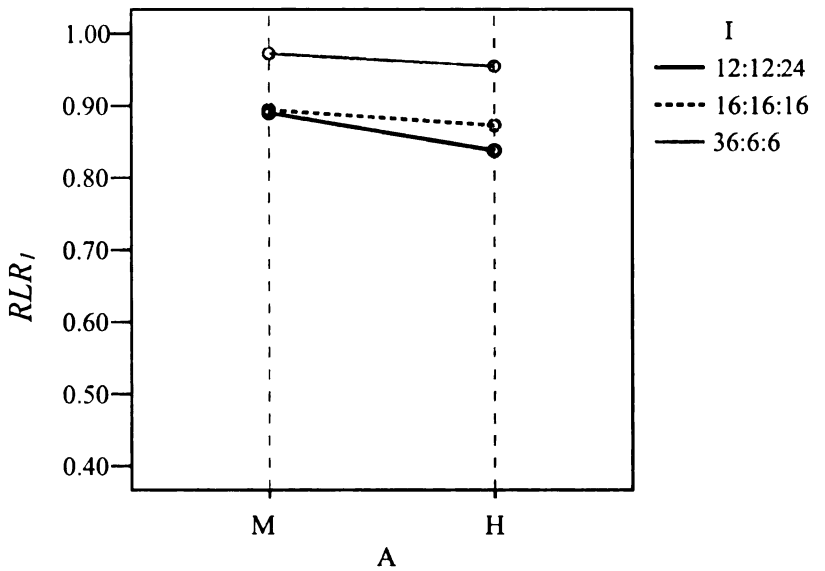


Figure 4.2.7. The interaction of A and I in RLR_1 given correlation matrix C_1

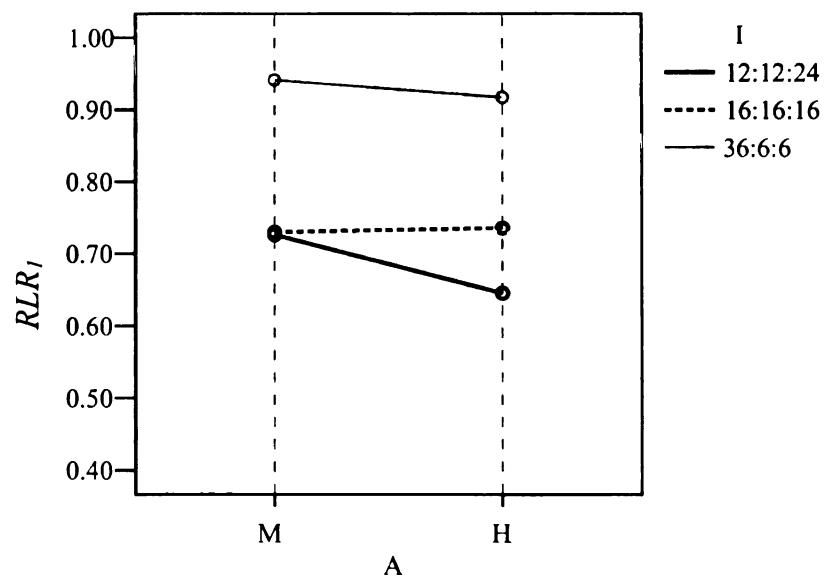


Figure 4.2.8. The interaction of A and I in RLR_1 given correlation matrix C_2

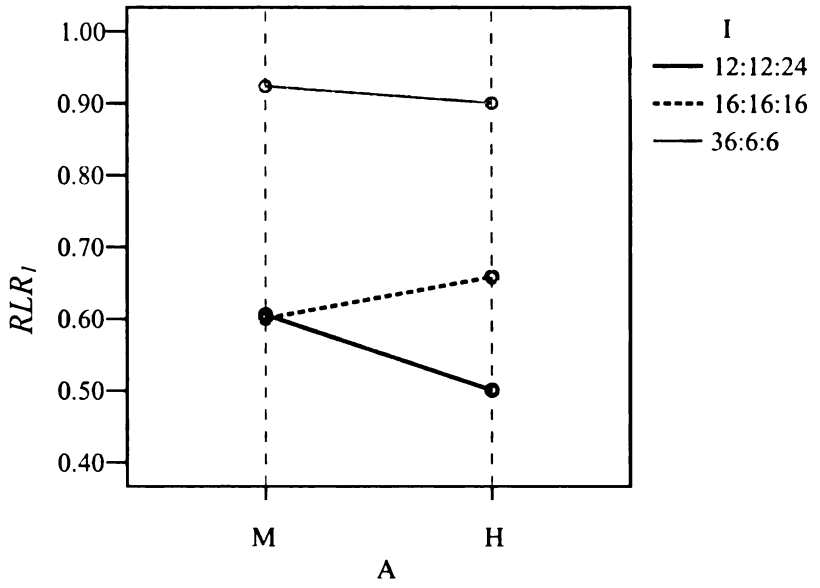


Figure 4.2.9. The interaction of A and I in RLR_1 given correlation matrix C_3

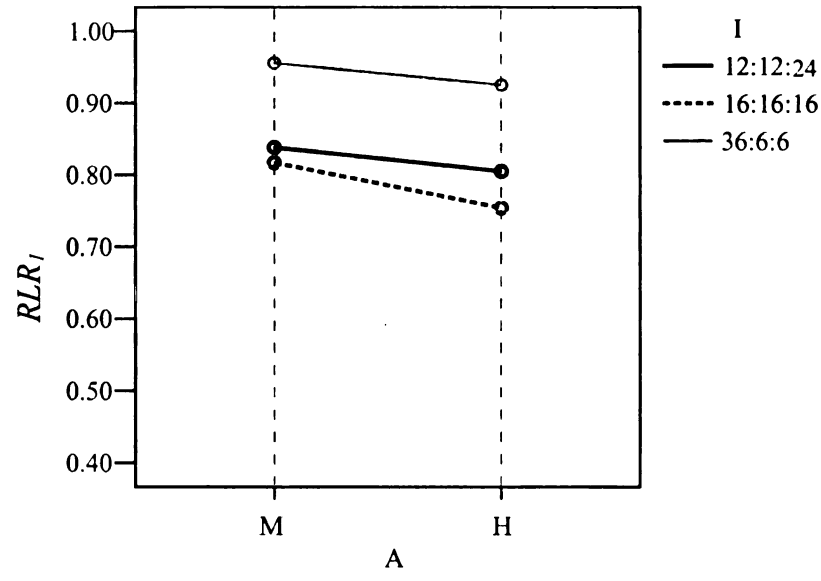


Figure 4.2.10. The interaction of A and I in RLR_1 given correlation matrix C_4

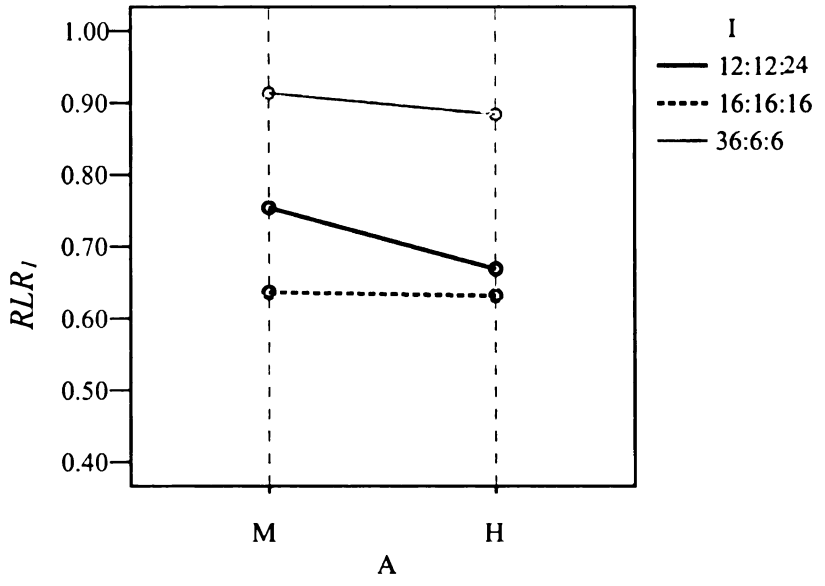


Figure 4.2.11. The interaction of A and I in RLR_1 given correlation matrix C_5

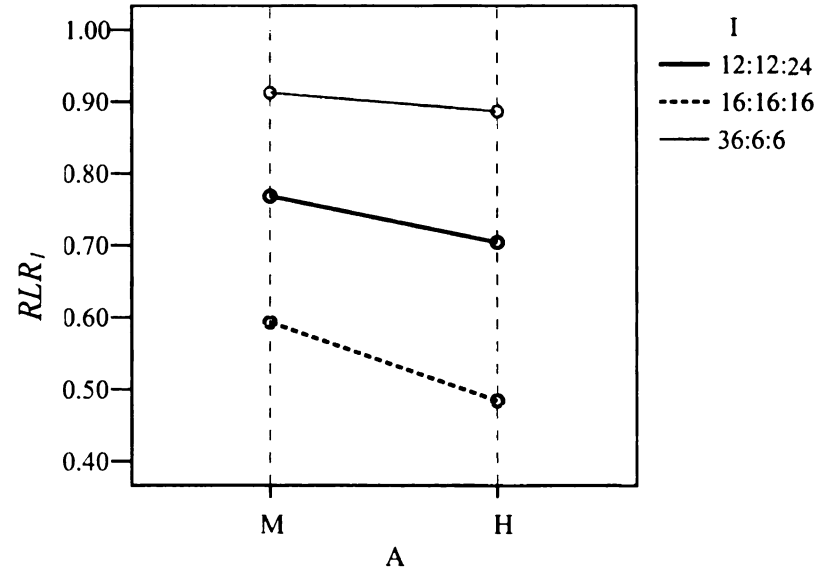


Figure 4.2.12. The interaction of A and I in RLR_1 given correlation matrix C_6

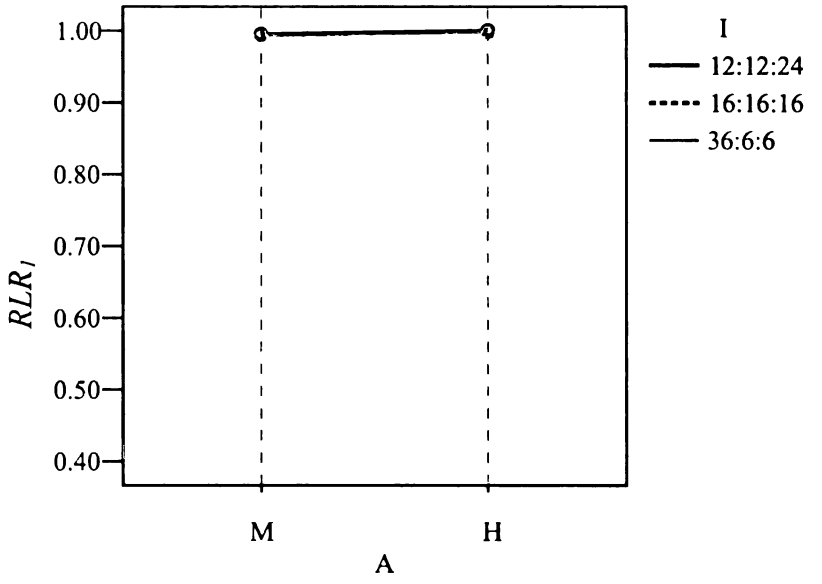


Figure 4.2.13. The interaction of A and I in RLR_2 given correlation matrix C_1

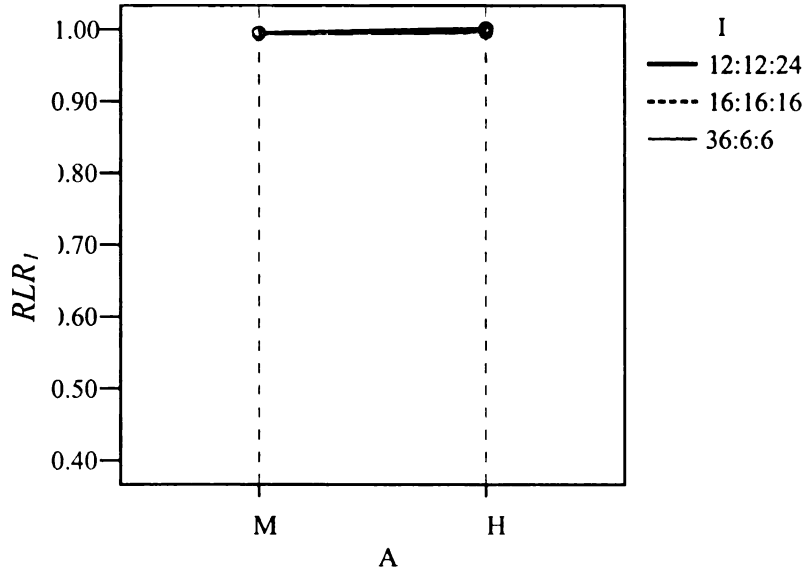


Figure 4.2.14. The interaction of A and I in RLR_2 given correlation matrix C_2

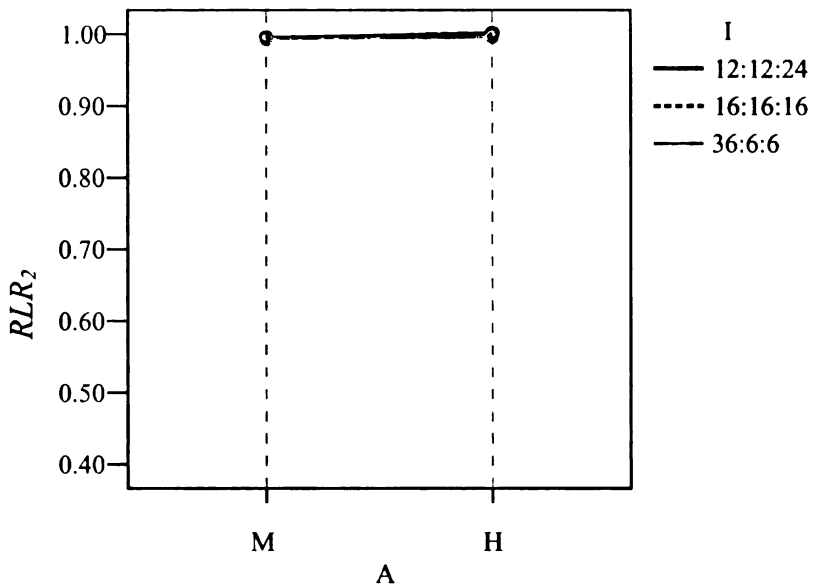


Figure 4.2.15. The interaction of A and I in RLR_2 given correlation matrix C_3

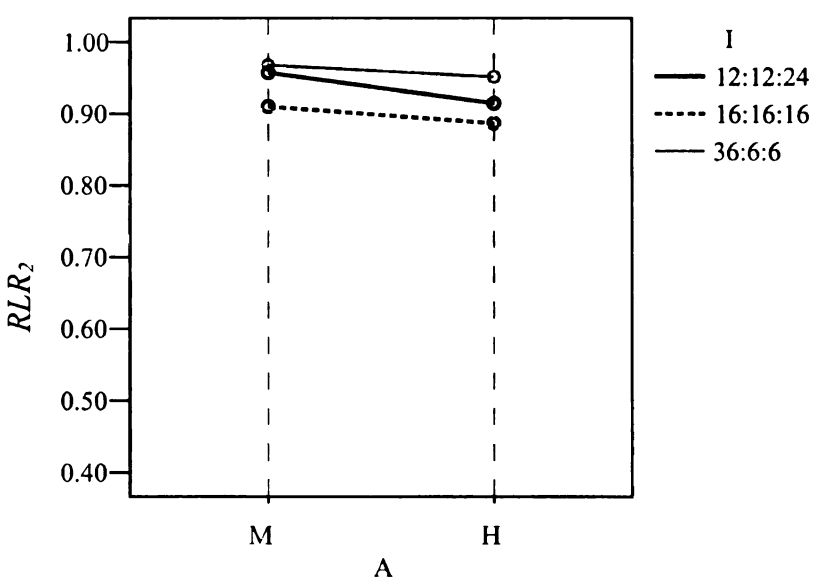


Figure 4.2.16. The interaction of A and I in RLR_2 given correlation matrix C_4

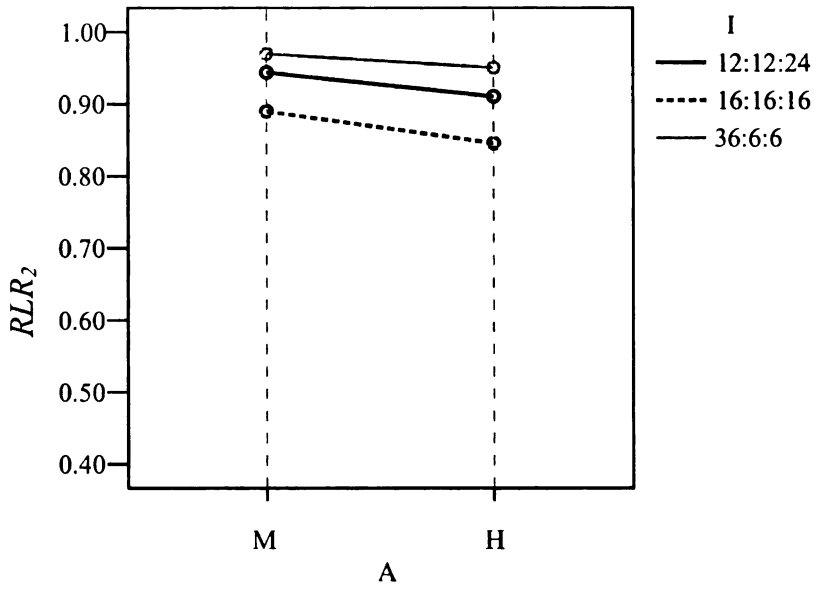


Figure 4.2.17. The interaction of A and I in RLR_2 given correlation matrix C_5

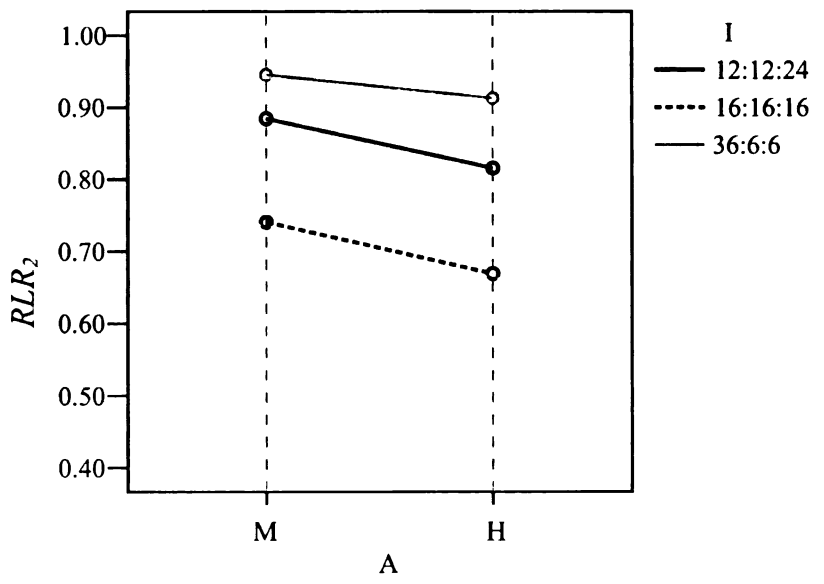


Figure 4.2.18. The interaction of A and I in RLR_2 given correlation matrix C_6

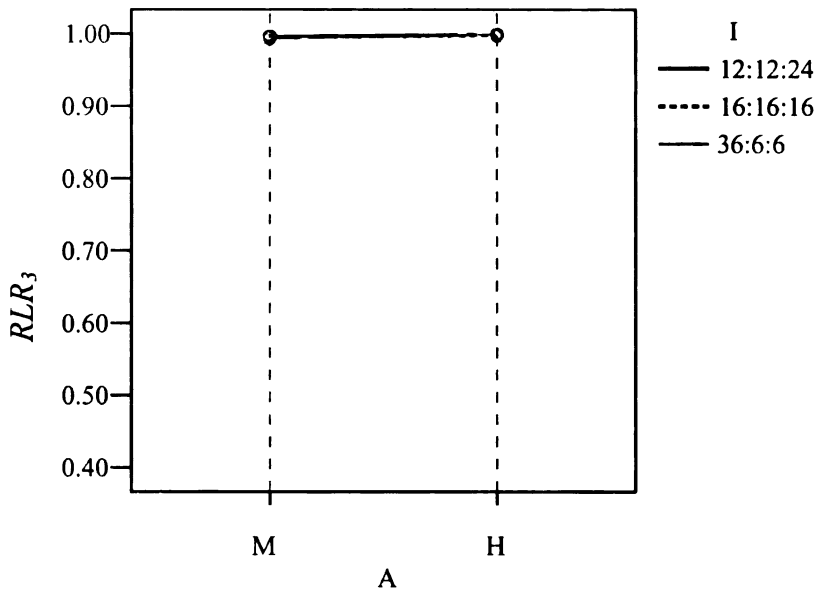


Figure 4.2.19. The interaction of A and I in RLR_3 given correlation matrix C_1

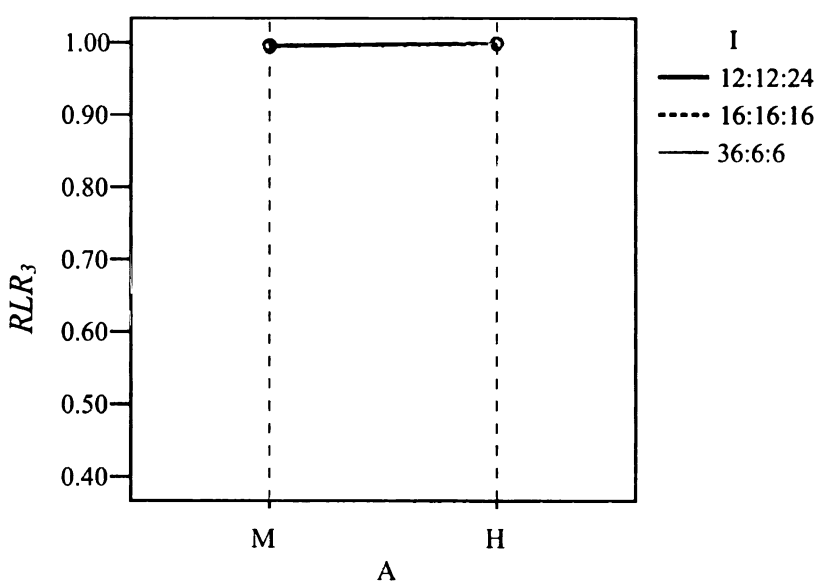


Figure 4.2.20. The interaction of A and I in RLR_3 given correlation matrix C_2

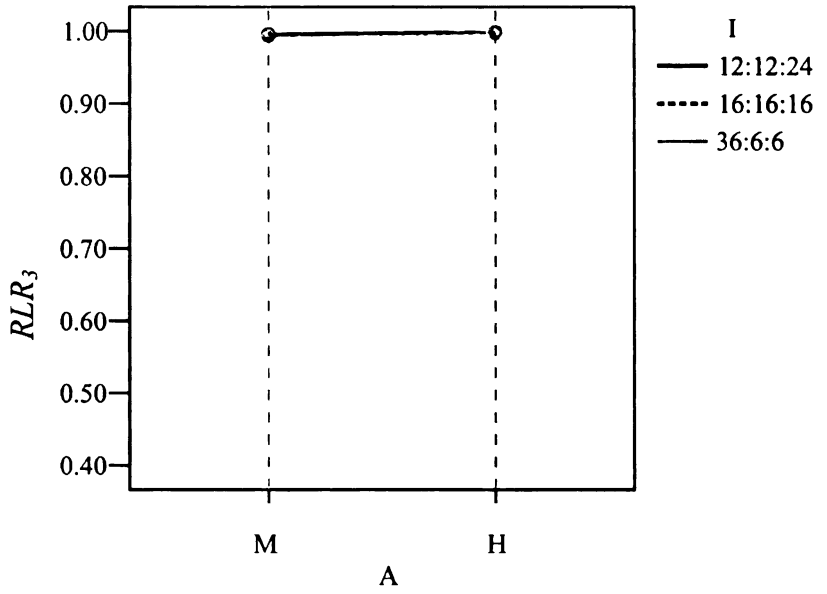


Figure 4.2.21. The interaction of A and I in RLR_3 given correlation matrix C_3

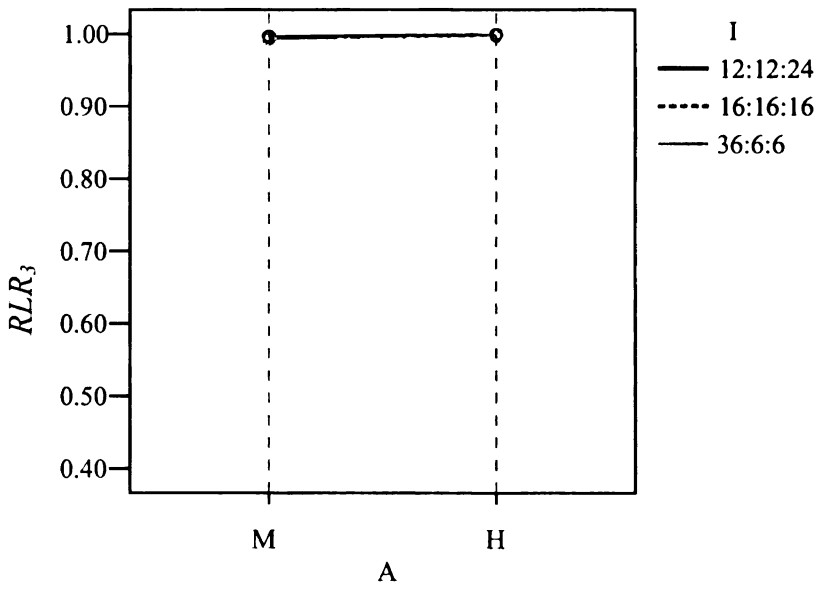


Figure 4.2.22. The interaction of A and I in RLR_3 given correlation matrix C_4

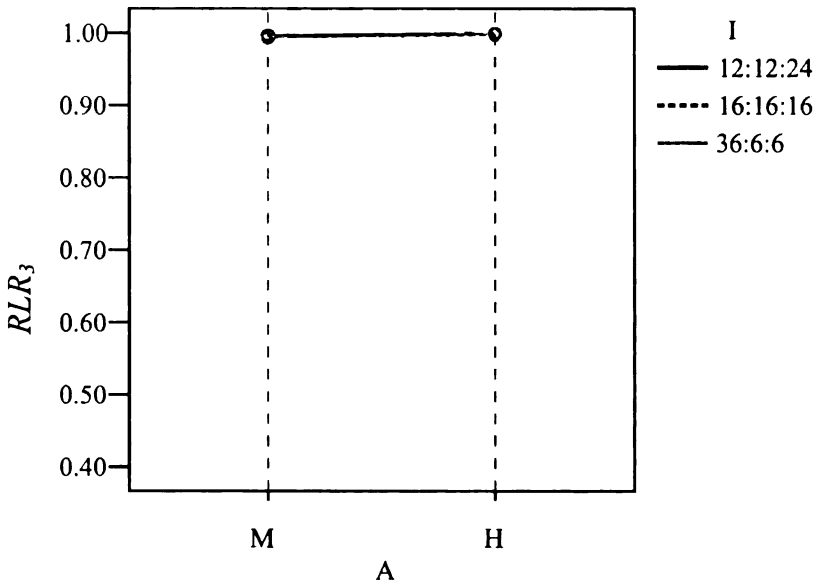


Figure 4.2.23. The interaction of A and I in RLR_3 given correlation matrix C_5

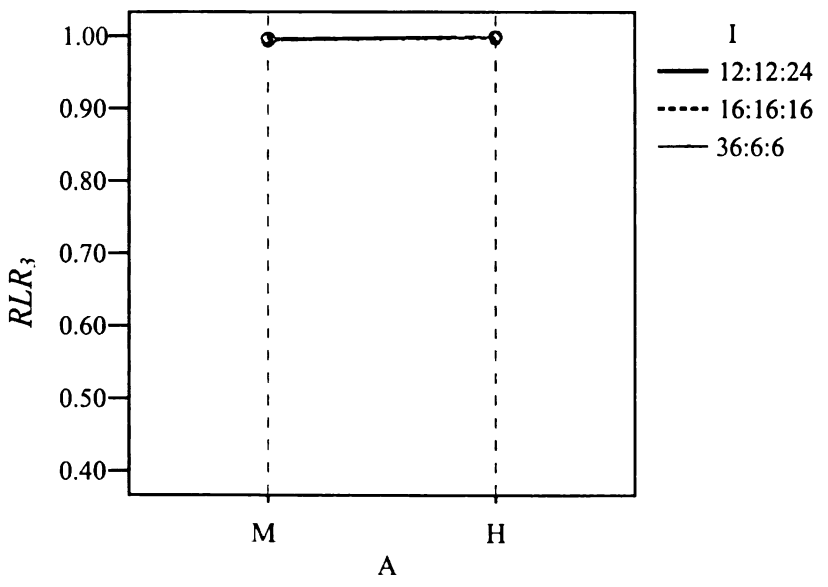


Figure 4.2.24. The interaction of A and I in RLR_3 given correlation matrix C_6

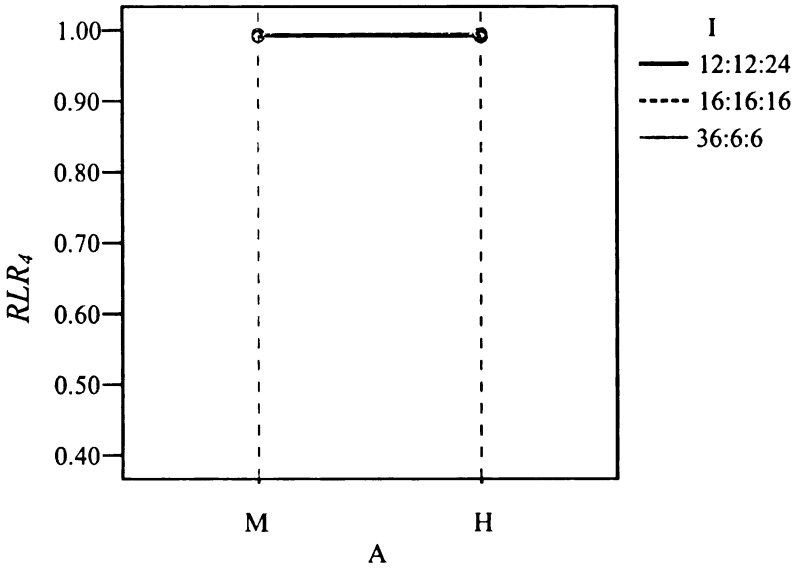


Figure 4.2.25. The interaction of A and I in RLR_4 given correlation matrix C_1

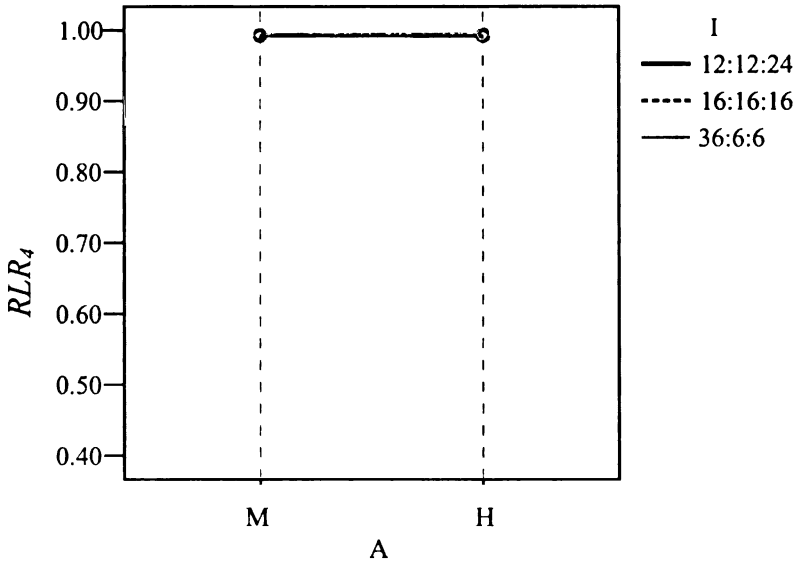


Figure 4.2.26. The interaction of A and I in RLR_4 given correlation matrix C_2

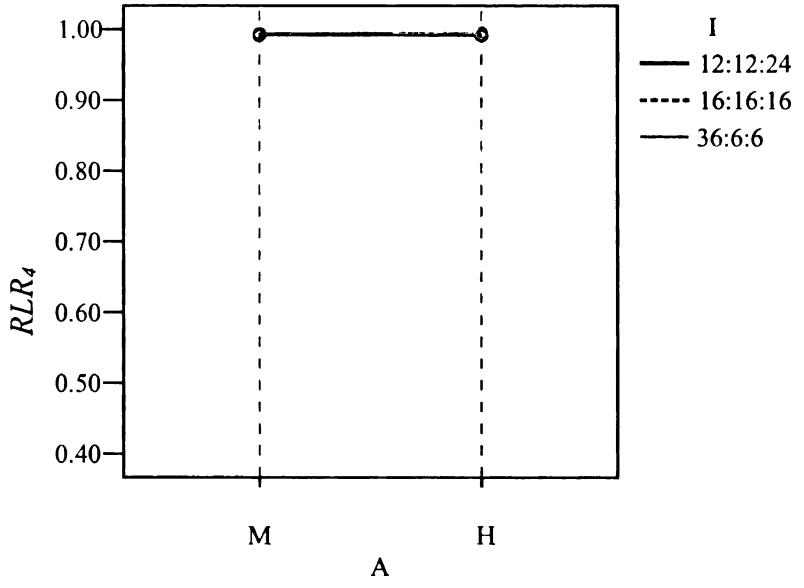


Figure 4.2.27. The interaction of A and I in RLR_4 given correlation matrix C_3

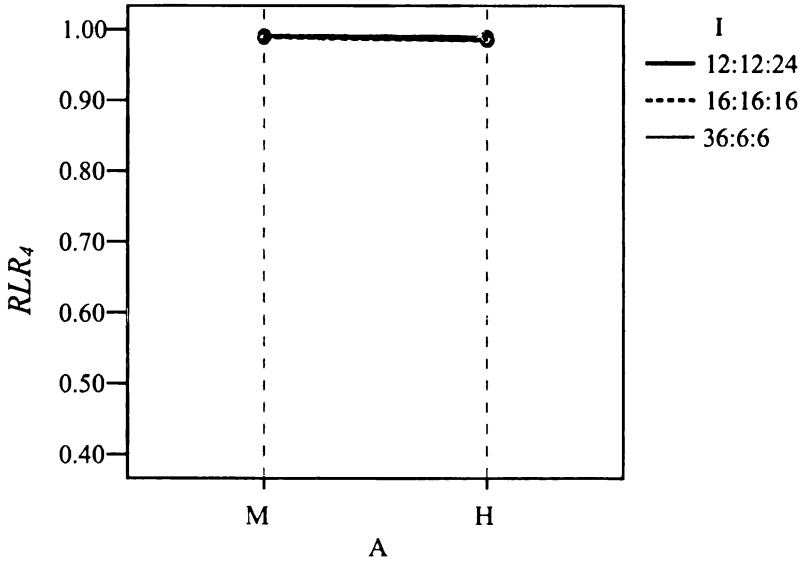


Figure 4.2.28. The interaction of A and I in RLR_4 given correlation matrix C_4

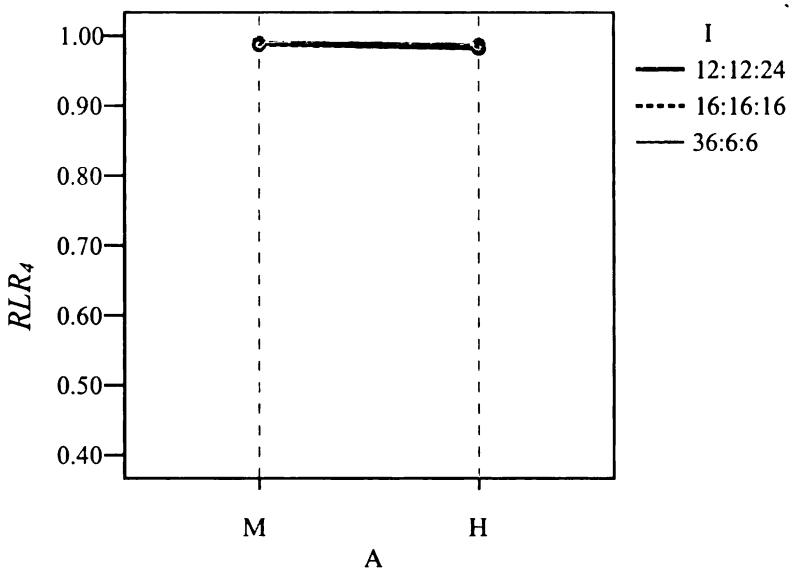


Figure 4.2.29. The interaction of A and I in RLR_4 given correlation matrix C_5

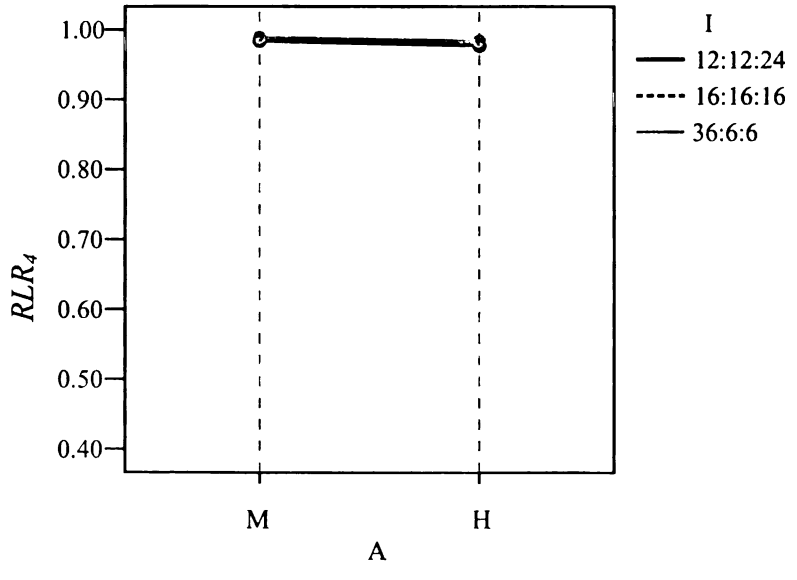


Figure 4.2.30. The interaction of A and I in RLR_4 given correlation matrix C_6

With regard to Figure 4.2.19 to Figure 4.2.24, all RLR_3 approached 1 regardless of the levels of *A*, *C*, and *I*. Because all the data sets were generated with a three-dimensional correlation matrix and three-dimensional item parameters, the model could fit the data well for any combination of *A*, *C*, and *I*. Concerning the results in Figure 4.2.25 to Figure 4.2.30, even though the RLR_4 were close to 1, discrepancies were found among RLR_4, especially when *A* was high. This explained that the effects of *A*, *C*, and *I* might still be important when the model over-fit the data.

4.2.3 Comparisons of the Numbers of Rejections

This section reports comparisons of the statistical power and the Type I error rate of the *RLR* index with those of the G^2 test and the G^2_{diff} test. Again, the theoretical α used for the G^2 test and the G^2_{diff} test was .05.

As shown in Table 4.2.10, given that the data were two-dimensional, the correct rejections of a unidimensional model were perfect with G^2 test and the G^2_{diff} test. Based on the unidimensional regression model built in Study I, the rejections based on the *RLR* index were correct except for Form 131. This finding indicated that the *RLR* index tended to underestimate the dimensionality for the two-dimensional data when the inter-factor correlation was as high as 0.7, item discriminations were moderate, and a weak minor factor sensitive to 6 items existed.

In order to test null hypothesis H_o: d= 2, the two-dimensional regression model was built. Given that H_o: d= 2 is true, the regression model was identified with the predictors of the estimated *a*-parameters (EA_1 and EA_2), the slope of the eigenvalues from

the estimated inter-factor correlation matrix (*ES*), and the estimated percentage of items having dominant loadings on the first and second dimension, respectively (PI_1 and PI_2). The overall model was significant with adjusted R^2 equal to .819 and can be expressed as

$$RLR_2 = 0.974667 + 0.036129(EA_1) + 0.001451(EA_2) - 0.004751(ES) + 0.011826(PI_1) - 0.027196(EA_1 \times PI_1) + 0.025021(EA_2 \times PI_2) + 0.008707(EA_1 \times ES) - 0.006682(EA_2 \times ES) + 0.004379(PI_1 \times ES) - 0.004752(EA_1 \times PI_1 \times ES) + 0.023656(EA_2 \times PI_2 \times ES). \quad (42)$$

If the observed RLR_2 fell in the 5% rejection area at the lower end of the distribution representing a good fit, the null hypothesis was rejected. The numbers of rejections of the two-dimensional model were also listed in Table 4.2.10. The *RLR* index generated false rejections less than 5 regardless of the levels of *A*, *C*, and *I*. On the contrary, the G^2 test and the G^2_{diff} test generated high rejections for all cases. Thus, for the two-dimensional data, the *RLR* index outperformed the G^2 test and the G^2_{diff} test by having low Type I error rates.

Table 4.2.10. The number of rejections in 100 replications for two-dimensional data

Data	H_0: d=1			H_0: d=2		
	RLR	G^2	G^2_{diff}	RLR	G^2	G^2_{diff}
Form 111	100	100	100	0	100	100
Form 112	100	100	100	1	100	100
Form 121	100	100	100	4	100	100
Form 122	100	100	100	0	100	100
Form 131	0	100	100	0	100	100
Form 132	100	100	100	0	100	100
Form 211	100	100	100	3	100	100
Form 212	100	100	100	5	100	100
Form 221	100	100	100	2	100	100
Form 222	100	100	100	0	100	100
Form 231	97	100	100	1	100	100
Form 232	100	100	100	0	100	100
Form 311	100	100	100	3	100	100
Form 312	100	100	100	4	100	100
Form 321	100	100	100	2	100	100
Form 322	100	100	100	1	100	100
Form 331	100	100	100	2	100	100
Form 332	100	100	100	2	100	100

For the three-dimensional data, similar procedures were used to decide the number of correct rejections of unidimensional and two-dimensional models. As Table 4.2.11 shows, with the theoretical α equal to .05, the G^2 test and the G^2_{diff} test perfectly rejected the wrong unidimensional and two-dimensional models and generated satisfactory statistical power. With regard to the *RLR* index, the rejection of a unidimensional model was based on the unidimensional regression model built in Study I and the results were satisfactory. The rejection of a two-dimensional model was based

on the two-dimensional regression model in equation (42) and the statistical power was perfect.

The three-dimensional regression model was built given that the null hypothesis H_o: d= 3 is true. With the predictors of the estimated a-parameters (EA_1, EA_2 and EA_3), the slope of the eigenvalues from the estimated inter-factor correlation matrix (ES), and the estimated percentage of items having dominant loadings on the first, second, and the third dimension, respectively (PI_1, PI_2, and PI_3), the three-dimensional regression model was built having adjusted R^2 equal to .384.

$$\begin{aligned} RLR_3 = {} & 0.035961 - 0.017669(EA_1) + 0.017799(EA_2) + 0.017707(EA_3) - 0.055493(PI_1) - \\ & 0.062399(PI_2) - 0.033739(EA_1 \times ES) - 0.002514(EA_2 \times ES) + 0.013083(EA_3 \times ES) + \\ & 0.038326(EA_1 \times PI_1) - 0.066626(EA_2 \times PI_2) - 0.061094(EA_3 \times PI_3) - \\ & 0.003045(PI_1 \times ES) - 0.046190(PI_2 \times ES) + 0.077303(PI_3 \times ES) + \\ & 0.057431(EA_1 \times PI_1 \times ES) - 0.028220(EA_2 \times PI_2 \times ES) - \\ & 0.090448(EA_3 \times PI_3 \times ES). \end{aligned} \quad (43)$$

If the observed RLR_3 was less than the lower bound of the distribution representing a good fit, the null hypothesis was rejected. The numbers of rejections of the three-dimensional model were listed in Table 4.2.11. Regardless of the levels of A, C, and I, the RLR index generated false rejections less than 5 times. On the contrary, the G^2 test and the G^2_{diff} test produced high rejections. For the three-dimensional data, the RLR index outperformed the G^2 test and the G^2_{diff} test by having low Type I error rates.

Table 4.2.11. The number of rejections in 100 replications for three-dimensional data

Data	H_0: d= 1			H_0: d= 2			H_0: d= 3		
	RLR	G^2	G^2_{diff}	RLR	G^2	G^2_{diff}	RLR	G^2	G^2_{diff}
Form 411	100	100	100	100	100	100	0	100	43
Form 412	100	100	100	100	100	100	0	100	53
Form 421	100	100	100	100	100	100	4	100	26
Form 422	100	100	100	100	100	100	0	100	57
Form 431	100	100	100	100	100	100	0	100	74
Form 432	100	100	100	100	100	100	0	100	46
Form 511	100	100	100	100	100	100	0	100	41
Form 512	100	100	100	100	100	100	2	100	66
Form 521	100	100	100	100	100	100	2	100	40
Form 522	100	100	100	100	100	100	0	100	52
Form 531	100	100	100	100	100	100	0	100	75
Form 532	100	100	100	100	100	100	0	100	48
Form 611	100	100	100	100	100	100	1	100	45
Form 612	100	100	100	100	100	100	4	100	92
Form 621	100	100	100	100	100	100	4	100	67
Form 622	100	100	100	100	100	100	5	100	93
Form 631	98	100	100	100	100	100	1	100	45
Form 632	100	100	98	100	100	100	0	100	55

4.3 Real Data Analysis

As a real data example, the Grade 4 Mathematics Test data from the MEAP testing program were employed. Instead of analyzing the whole data set, five independent random samples of 2000 examinees were randomly selected. For each level of dimensionality, the *RLR* index was calculated and listed in Table 4.3.1. The results indicated that the values of RLR_l were as high as .97 in all five samples. When adding dimensions to the model, all the *RLR* values didn't approach 1. This pattern of the *RLR* values approximated the results in the unidimensional simulation. With a sample size of

2000, when the data were truly unidimensional with moderate level of item discrimination, the *RLR* index stayed at a fixed level regardless how many dimensions were added to the model.

Table 4.3.1. The *RLR* indices for the MEAP Grade 4 Mathematics Test data

Sample	RLR_1	RLR_2	RLR_3	RLR_4
Sample 1	0.9828	0.9850	0.9827	0.9828
Sample 2	0.9798	0.9825	0.9801	0.9798
Sample 3	0.9748	0.9808	0.9784	0.9748
Sample 4	0.9713	0.9814	0.9791	0.9713
Sample 5	0.9749	0.9796	0.9787	0.9749

To decide whether or not the Grade 4 Mathematics Test data were unidimensional, the mean of the estimated *a*-parameters, the *SD* of estimated *d*-parameters, along with the sample size and test length were implemented in equation (41) to decide the lower bound of a good fit. Table 4.3.2 shows the descriptive statistics of item parameters and the lower bound of the predicted RLR_1 value. Because the null hypothesis H_0: d= 1 was not rejected, the significance test stopped at the unidimensional model. All the results indicated that this Mathematics Test data can be well fit by the unidimensional model.

Table 4.3.2. Item parameter estimates and the test of unidimensionality

Sample	Item parameter				Lower bound	H_0: d= 1
	Mean (a)	*SD*(a)	Mean (d)	*SD* (d)		
Sample 1	0.6662	0.2357	0.9151	0.6510	0.9561	Not rejected
Sample 2	0.6686	0.2210	0.9201	0.6540	0.9565	Not rejected
Sample 3	0.6481	0.2340	0.8757	0.6498	0.9531	Not rejected
Sample 4	0.6635	0.2274	0.8752	0.6432	0.9557	Not rejected
Sample 5	0.6574	0.2202	0.9042	0.6343	0.9547	Not rejected

CHAPTER 5

SUMMARY, DISCUSSION, AND CONCLUSION

In this chapter, the overall results are summarized and the related issues are discussed. The conclusions are drawn from both the simulation studies and the real data analysis. Moreover, the limitations of the research and suggestions for future studies are provided.

5.1 Summary of the Research

The purpose of this research was to propose a new index to evaluate the model-data-fit for the compensatory logistic MIRT model. Once the number of dimensions is identified to adequately describe the item response data, the item and ability parameters can be correctly estimated. Then, the test scores can be correctly estimated by the MIRT model and the subsequent testing techniques, such as test equating for multidimensional abilities, can possibly be conducted.

The *RLR* index proposed in this study is derived from Estralla's (1998) R^2 analog which is equivalent to the R^2 in the OLS model. The *RLR* index compares the percentages of the unexplained variance in the k-dimensional MIRT model with that in the $(k+1)$-dimensional MIRT model. The value of the *RLR* index reflects the improvement of fit obtained by adding one more dimension to the MIRT model. When the model fits the data, the error reduction due to adding one more dimension to the model is limited and the *RLR* index approaches 1.

This research investigated the performance of the *RLR* index with respect to its

ability to correctly identify the dimensionality for both unidimensional and multidimensional data reflecting different levels of item discrimination, item difficulty, sample size, test length, inter-factor correlation, and item-factor structure. The statistical characteristics of the *RLR* index were compared to those of the G^2 test and the G^2_{diff} test. The test data from the MEAP Grade 4 Mathematics Test were analyzed to show how the *RLR* index decided the dimensionality of real data.

5.2 Discussion

Based on the results in Chapter 4, the major findings are highlighted in the following sections.

The Rates of Unsuccessful TESTFACT Runs

When unidimensional data were analyzed by MIRT models, some analyses were unsuccessful. Because the data generated a singular tetrachoric correlation matrix, the full-information factor analysis procedure stopped. When the tests were short, all the TESTFACT runs were successful. However, when the tests were long, the tests with high variation of the *d*-parameters generated a singular correlation matrix at higher rates than the tests with low variation of the *d*-parameters. This was more severe when the sample size was small. When the test was long, the size of the frequency table for calculating the pair-wise tetrachoric correlation was large, resulting in some cell frequencies being too small to give meaningful tetrachoric correlation estimates. For those invalid item-pairs, TESTFACT automatically used the substitute values of either 1 or -1. When the test was long but the sample size was small, the number of invalid

item-pairs increased and caused more inaccurate tetrachoric correlation estimates. Thus, with the limited number of valid item-pairs, the resulting tetrachoric correlation matrix tended to be problematic.

Given that the test was long, the rate of getting a singular matrix was greater when the *d*-parameters had higher variation. This finding is consistent with Roznowski et al's study (1991). The tetrachoric correlation has the special property that when it approaches either 0 or 1, the variation of the sampling distribution is large. In this study, when the test items were extremely easy or difficult, the underlying correlations for these item-pairs approached 0 or 1. Accordingly, these pair-wise correlations were poorly estimated and resulted in many arbitrary 0's and 1's. Again, with a large number of inaccurate tetrachoric correlation estimates, the tetrachoric correlation matrix would have a high probability to be singular, causing problems in full-information factor analysis.

For the multidimensional data, given the same combination of inter-factor correlation and item-factor structure, the data with moderate item discriminations generated a singular tetrachoric correlation matrix at higher reats than the data with high item discriminations. When the levels of inter-factor correlation and item discrimination were held constant, the third level (36: 6: 6) of the item-factor structure generated a singular correlation matrix at lower rates than the first level (12: 12: 24) and the second level (16: 16: 16) of item-factor structure. However, the inconsistent patterns were found between the two- and three-dimensional data sets. The two-dimensional data showed higher rates of getting a singular matrix for the four-dimensional model than for the five-dimensional model. Conversely, the three-dimensional data demonstrated lower rates of getting a singular matrix for the

four-dimensional model than the five-dimensional model. In order to explore this problem, different TESTFACT settings, such as different numbers of quadrature points in the EM algorithm, different levels of iteration cycles, and different levels of the convergence criteria were employed, but similar results were obtained. It is unclear how full-information factor analysis generated the inconsistent results. The performance of TESTFACT computer program needs further investigation in future studies.

The Unexpected Values of the RLR Index and the G^2_{diff} Test

As shown in the summary statistics of Study II, unexpected *RLR* values (*RLR*>1) were found in the multidimensional simulation. These unexpected values occurred when the estimation model recovered the true dimensionality or over-fit the data. Theoretically, the *RLR* index should not be greater than 1 because the SSE should not increase when adding more factors to the model. However, whenever the *RLR* index exceeded 1, the corresponding G^2_{diff} test generated a negative value, which was not reasonable for a χ^2 distribution. The exact cause of these unexpected values was not clear, but a possible explanation is provided.

The R^2 for the OLS model has the property of not decreasing when more predictors are added to the model. However, this is not always the case for the MIRT model where both the *a*-parameters and ability parameters need to be estimated simultaneously. Adding one more factor to the MIRT model increases the degrees of freedom, but simultaneously requires $m + n - 2$ (n is the number of items, and m is the number of examinees) more parameters to be estimated. It is possible that, when the model-data-fit

is already perfect, adding more factors to the MIRT model would increase fit, but simultaneously would generate larger estimation errors. When the model over-fits the data, the increase of fit due to adding more factors may not compensate for the increase of estimation error. According to the definition, the *RLR* index is the ratio of the log transformation of the unexplained percentage of the variance from the *k*-dimensional and the (*k*+1)-th dimensional models. When the unexplained percentage of the variance of the *k*-dimensional model is smaller than that of the (*k*+1)-th dimensional model, the value of the *RLR* index becomes greater than one.

The same rationale can be applied to explain the negative values of the G^2_{diff} test. The G^2 test in equation (13) is a discrepancy function based on the ratio of the likelihood under the fitted model to the likelihood of the empirical frequencies. The G^2_{diff} test, as shown in equation (14), compares the discrepancy of the likelihoods for the model and the data between a lower-factor model and a successive higher-factor model. The formula explicitly indicates that the discrepancy between the model and the data for the lower-factor model should always be greater than the discrepancy between the model and the data for the higher-factor model. In this study, however, the results showed that the assumption of the formula is not always true. When the model already fits the data well, over-fitting the data by adding one more factor to the model may increase the discrepancy between the model and the data and thus generates a negative value for the G^2_{diff} statistic. Because both the *RLR* index and the G^2_{diff} test compare the fit of the two successive MIRT models, the over-fitting problem occurs when the lower-factor model already has a good fit and the higher-factor model has a relatively

poor fit. Thus, when the over-fitting problem arises, the *RLR* index may exceed 1, and at the same time the G^2_{diff} statistic may be negative.

The Patterns of the RLR Values and Dimensionality

From the results in unidimensional simulation, RLR_1 reached .99 when the *a*-parameters were higher than 0.2. In such cases, because all RLR_1 values were close to the upper bound, adding more dimensions to the model only increased the values of RLR_2 and RLR_3 at the third decimal place. Conversely, for the tests with the *a*-parameters equal to 0.2, adding factors to the model did obviously increase the *RLR* values.

The simulation of two-dimensional data based on the three-dimensional inter-factor correlations and the three-dimensional item parameters was successful. The patterns of the *RLR* values for the multidimensional data sets, as shown from Figure 4.2.1 to Figure 4.2.6, were as expected. For the two-dimensional data, the values of RLR_1 were small, but the values of RLR_2 approached 1. When adding more factors to the model, both the values of RLR_3 and RLR_4 were still close to 1. For the three-dimensional data, the values of RLR_1 were small. When adding a second factor to the model, the values of RLR_2 increased but not to the level of a good fit. For the three-dimensional solution, all the values of RLR_3 approached 1, suggesting a good fit. When the model over-fit the data, the values of RLR_4 were still close to 1, but sometimes less than the values of RLR_3.

Based on the results of the unidimensional and multidimensional simulation studies, it was clear that the change of the *RLR* values with dimensionality reflected the simulated dimensionality underlying the data. Once the *RLR* index stops increasing, the minimum number of statistical dimensions can be specified.

The Variables Associated with the RLR Index and Dimensionality

The results from the MANOVA analysis in Study I showed that item discrimination, item difficulty, sample size, and test length collectively had an effect on the *RLR* index. Sample size affected the *RLR* index, but the effect varied depending on the level of item discrimination. A large sample size helped reducing the sampling variation and offered better estimates of the model parameters, especially when the item discrimination was low. With a larger sample size, the *RLR* index became more stable. That is, when item discrimination was low, the problem of falsely rejecting the true unidimensionality was circumvented. The effect of item difficulty also depended on the level of item discrimination. As long as item discrimination was greater than 0.2, the effect of item difficulty was minor.

The results based on the MANOVA analysis in Study II indicated that inter-factor correlation, item-factor structure, and item discrimination all together influenced the *RLR* index. Because the interactions were significant and some of them had substantive magnitude of effect sizes, the simple effect instead of the main effect should be discussed. Given the same level of inter-factor correlation and item-factor structure, high item discrimination increased the change of *RLR* associated with dimensionality when the model under-fit the data. Thus, the judgment of the dimensionality based on the *RLR* index would be easy when the item discrimination was high. Given the same level of inter-factor correlation and item discrimination, the change of *RLR* with dimensionality was the greatest when items were evenly sensative to factors. In other words, when there was no clear dominant factor in the data, the change of *RLR* with dimensionality

was obvious. On the contrary, when the data had a strong dominant factor and some weak minor factors that were only sensitive to a small number of items, the change of *RLR* with dimensionality became small and thus increased the difficulty of identifying minor factors. However, when the model fit the data, the effects of item discrimination, inter-factor correlation, and item-factor structure became minor.

The RLR Index and the Magnitude of the Dominant Factor

In terms of the factor analysis technique, the dominant factor will always be identified first by the factor-analytical model. Then, minor factors will be extracted in order by their quantities of explained variance. The first extracted factor always explains the most variance in the data than the subsequent factors. The R^2 technique is primarily designed to represent the percentage of explained variance in the data. In the MIRT model, R_1^2 shows the percentage of variance explained by the unidimensional model, and R_2^2 reflects the percentage of variance explained by the two-dimensional model. Based on the equivalence between the MIRT model and factor analytic model, RLR_1 can be used to show the relative size of the dominant factor in contrast to the second factor.

Based on the results from the unidimensional simulation, it is clear that the magnitude of RLR_1 was related to the size of item discrimination. RLR_1 reached .99 when item discrimination was 0.4 or higher. Even though item discrimination was as low as 0.2 with a short test and a small sample size, the minimum value of RLR_1 was .80. For the unidimensional data with higher item discrimination, the dominant factor explained more variance in the data and thus could be more easily identified by the

statistical model.

The determination of the size of the dominant factor is more complex in the multidimensional data. When inter-factor correlation and item discrimination were held constant, RLR_1 increased with the increment of the number of items sensitive to the dominant factor. The two-dimensional data, the data related to the correlation matrices C_1, C_2, and C_3, were generated with a three-dimensional inter-factor correlation matrix and item-factor structure by combining the first two groups of items into a bigger item cluster. Thus, the first level of the item-factor structure (12:12:24) generated the lowest dominant factor, which were sensitive to 50% of items in a test. The second level of the item-factor structure (16:16:16) produced a dominant factor sensitive to 67% of items in a test. With 88% of items sensitive to one factor, the third level of the item-factor structure (36:6:6) generated the greatest value of the dominant factor and at the same time had the greatest value of RLR_1. With regard to the three-dimensional data, which were the data sets related to C_4, C_5, and C_6, the percentage of items related to one factor was consistent with the level of item-factor structure. For the second level of the item-factor structure (16:16:16), each of the three dimensions had 33% of items. Without a doubt, the second level of item-factor structure (16:16:16) generated lower RLR_1 than the first level (12:12:24) and the third level (36:6:6) of item-factor structure. With 76% of items related to the main factor, the third level of the item-factor structure had the largest dominant factor and generated the greatest value of RLR_1.

Given the same level of item-factor structure and item discrimination, RLR_1 increased proportionally to the decrease of the inter-factor correlations. In factor analysis, when the factors are completely independent, the dominant factor tends to

explain less variance than the case when the factors are correlated. Thus, it is not surprising that when the level of item-factor structure and item discrimination were held constant, C_3 generated the lowest value of RLR_I in the two-dimensional data and C_6 generated the lowest value of RLR_I in the three-dimensional data. In short, RLR_I in the multidimensional data reflected the size of the dominant factor. Low RLR_I suggested that the items were more evenly distributed to factors, and the factors tended to be independent of each other. Correspondingly, the lower value of RLR_I also implied that the data were less likely to be unidimensional.

The Statistical Characteristics of the RLR Index, G^2 Test, and G^2_{diff} Test

The results of the G^2 test and G^2_{diff} test indicated that these statistics could not accurately identify dimensionality. Even though these statistics demonstrated high statistical power in rejecting wrong models, they tended to reject right models with high Type I error rates. These findings are consistent with earlier studies (Berger & Knol, 1990; De Champlain & Gessaroli, 1998; DeMars, 2003; McDonald, 1989b) that these G^2 tests should not be used to assess the dimensionality for test data. On the contrary, the *RLR* index demonstrated low Type I error rates and high statistical power for most data sets.

In the unidimensional simulation, the *RLR* index generated low Type I error rates except for the extreme cases when item discrimination was 0.2 and sample size was 2000. When item discrimination is low and sample size is limited, the test data are close to random data so that the signal in the data is unnoticeable. Accordingly, it is reasonable

that the *RLR* index can not function well for these test data. From the practical consideration in test development, a test with these items can be considered useless because items are not discriminating examinees' abilities. It can be expected that such bad tests may not be developed in real testing conditions, so the failure of the *RLR* index in detecting the true unidimensionality for these test data will not be an issue. It can be concluded that the *RLR* index demonstrated low Type I error rates for common tests. When the data are close to random, the index tended to falsely reject the true unidimensional model.

With regard to the multidimensional data, the *RLR* index performed well in rejecting the wrong unidimensional model except for the two-dimensional data having two highly correlated factors, a strong dominant factor, and moderate item discrimination. For this kind of test data, the *RLR* index cannot detect the weak second factor and tends to underestimate the data dimensionality. Other than this special case, the *RLR* index had high statistical power and low Type I error rates. The results of the simulation studies indicated that the *RLR* index outperforms the G^2 test and the G^2_{diff} test in detecting the true dimensionality.

Real Data Analysis

The *RLR* indices for the five random samples consistently indicated that the Grade 4 Mathematic Test data from the MEAP testing program can be modeled unidimensionally. As described earlier, this test was designed to measure different ability domains and skills in mathematics at the grade-4 level. The results based on the *RLR* index suggested that these content domains may be described under the umbrella of a general factor called

"basic mathematics skills." The unidimensional finding is supplemented with the discussions in term of the test item content, the representativeness of the content-related dimension, the definition of dimensionality, and the assumption of the compensatory MIRT model.

The mathematics knowledge taught in grade-4 contains the basic mathematics concepts and skills. The differences among different content knowledge and skills may not be as great as expected by the test developers. For example, if students can do multiplication, they need to have the prerequisite knowledge in addition. When responding to fraction questions, students have to think about how fractions are related to a unit whole, compare fractional parts of a whole, and find equivalent fractions to give a correct response. The processes for answering these mathematics questions are actually related to counting and addition. As a whole, the test items in the Grade 4 Mathematics Test may cover several distinct content domains, but these content-related abilities may be indeed highly correlated to each other. As shown in the second simulation study, when two of the three factors are highly correlated, the dimensions will converge so that a two-dimensional model can well explain the truly three-dimensional data. When the content-related abilities are highly correlated, similar to the multicollinearity problem in multiple regression, it is difficult to identify the net contribution of the minor factors when the dominant factor already explains most of the contribution of the minor factors.

Besides, how well the minor factors were measured in the Mathematics Test is another important issue. The Mathematics Test contained 57 items: 6 items for data and probability; 6 items for geometry; 18 items for measurement; and 27 items for numbers and operations. For the 6 items in data and probability, the mean of the item

discrimination is only 0.5347. With regard to the 6 items in geometry, the mean of the item discrimination is 0.5413. Given that the content-related abilities are highly correlated, those weak dimensions having only 6 moderate-discriminating items are not easily identified by a mathematical model.

Another explanation for the findings from the real data analysis goes back to the definition of dimensionality. There appears to be a common misconception that a set of items on a test measure a distinct number of dimensions regardless of the characteristics of the examinees taking the test (R. L. Turner et al., 1996). However, the statistical dimensionality is a characteristic of the data matrix, not the test or examinee population (Reckase, 1990). Researchers (Ackerman, 1994; Reckase, 1997a; R. L. Turner et al., 1996) pointed out that dimensionality is a function of both the skills being measured by the items and the multivariate ability distributions of the examinees. The dimensional structure of the data from a test could differ for various subgroups of an examinee population. Ackerman (1994) indicated that if items collectively are capable of distinguishing between levels of several skills, and examinees differ in their levels of proficiencies on more than one of these skills, the interaction needs to be described by a multidimensional model. Based on this rationale, the findings of the Grade 4 MEAP Mathematics Test data may indicate that these test items indeed covered several distinct content domains and the items should be described by more than one content-related ability, but the target examinees, i.e. the grade-4 students in Michigan state, were heterogeneous with respect to the main content-related ability but homogeneous with respect to the minor content-related abilities. When the variations of examinees' proficiencies on the minor content-related abilities were limited, it is difficult for a

mathematical model to capture those dimensions.

Another possible explanation can be offered based on the assumption of the compensatory logistic MIRT model. This MIRT model assumes that abilities can be linearly combined and compensated. It is possible that the content-related dimensions for the Mathematics Test data may be multidimensional, but the items were sensitive to the same combination of the content-related dimensions. Consequently, the statistical dimension needed for the model to describe the item-person interaction was one. Given the unclear nature of the ability structure in the mathematics test data, it is uncertain whether or not the unidimensional model can still fit the data well if a different model, such as a partially compensatory model, is used to analyze the same data.

To conclude these possible explanations for the real data analysis, one statistical dimension was enough to sufficiently explain the MEAP Grade 4 Mathematics Test data when the compensatory logistic MIRT model was used.

5.3 Conclusion

Based on the findings in the simulation studies and the real data analysis, the *RLR* index is a promising goodness-of-fit index for the MIRT model. The dimensionality index varied in accuracy as a function of sample size and could more accurately identify unidimensionality as the number of items increased. The *RLR* index demonstrated low Type I error rates except for the tests composed of poor items having item discrimination values of 0.2 with a short test and a small sample size. The *RLR* index also revealed high statistical power in rejecting wrong models except for the two-dimensional data with highly correlated factors, moderate item discrimination, and one weak minor factor.

The change of the *RLR* index with dimensionality implied the decrease of error in the data when adding factors to the model. Moreover, the *RLR* index for the initial unidimensional model reflected the size of the dominant factor. When the *RLR* index for the initial unidimensional model was low, it implied that the data had a weak dominant factor and were less likely to be unidimensional. Based on the *RLR* index, the Grade 4 Mathematics Test data from the MEAP testing program can be well explained by the unidimensional model. Even though the test was developed by selecting items representing different knowledge domain and skills, one statistical dimension would be enough to explain the interaction between items and examinees.

5.4 Limitations, Implications, and Suggestions for Future Research

The purpose of this study is to offer an index which can be used as a rule of thumb in selecting the most appropriate dimensionality for the MIRT model to explain test data. Instead of relying on subjective judgments, the proposed index provides objective and useful information to decide dimensionality based on the compensatory logistic MIRT model. Once the dimensionality is identified, the dimensional structure can further be explored to identify the relationships between dimensions. Validity studies (to identify what domains or dimensions are measured) can proceed to provide evidence supporting hypothesized multidimensionality and to identify construct-irrelevant variance.

It is important to emphasize that these findings were just preliminary and caution should be taken when interpreting and generalizing the results to other conditions. It is therefore important to highlight the limitations associated with this investigation and to offer suggestions for future research with reference to assessing MIRT goodness-of-fit.

First, as introduced in Chapter 2, the parametric MIRT models provide full dimensionality estimation specifying the number of dimensions and which item measures which dimension, but these benefits all rest on their specific assumptions of the item responses. Tate (2002) pointed out that any mathematical model with limited numbers of parameters provided a relatively efficient summary of data, but it also brought in the strong assumption that the phenomenon of interest could be accurately explained by the assumed model. Based on the rationale, data dimensionality can be determined by the model-data-fit procedure only when the proposed model is appropriate. Since the *RLR* index was derived from the logistic compensatory MIRT model (Reckase, 1985; Reckase & McKinley, 1991), this index can work well only when the logistic compensatory MIRT model is the appropriate model to explain the data.

The logistic compensatory MIRT model used in this study is only one of the MIRT models proposed in the literature. This model explicitly assumes that abilities can be linearly combined so that the high level of one ability can compensate for the low level of a second ability. However, for real test data it is unclear if abilities can be linearly combined or compensated. Sympson's model, for example, assumes that the ability structure underlying the test data is partially compensatory (cited from Reckase & McKinley, 1982). A correct item response requires examinees to demonstrate high abilities on all dimensions. If the underlying dimensional structure in the data is different from the model assumption, using the model to explain the data may not generate a good fit unless the extremely high-dimensional model is used. As explained by Tate (2002), the attempt to fit the partially compensatory function with a compensatory model is similar to the unwise attempt to use an additive regression model

to represent an interactive relationship. However, so far the robustness of the compensatory MIRT models to the violation of the assumption of ability compensation is still unclear. It would be worth noting that the MIRT model used in this study is only one option to describe test data. If the inherent ability dimensions in the data cannot match the model assumption, using the compensatory MIRT model to describe the data may result in essential misfit, and consequently the statistical power of the *RLR* index would be limited.

Second, since the *RLR* index compares the ratio of the residuals of the two successive MIRT models, the degrees of freedom for the *RLR* index need further investigation. In the OLS model, R^2 is not an unbiased estimate of the corresponding parameter in the population, and the degree of bias depends on the relative size of the number of observations (N) and the number of parameters (P)(Howell, 2001, p. 546). In the OLS regression model, the number of parameters is usually independent of the number of observations. The R^2 tends to be perfect (R^2= 1) when N= P + 1 regardless of the true relationship between the dependent variable and the predictors in the population. For the MIRT models, however, the total number of parameters needed to be estimated is always large. As the number of examinees increases in the MIRT model, the number of parameters increases proportionally. For example, in a unidimensional MIRT model, if 2000 examinees take a test that has 40 items, the total number of parameters to be estimated is 2078 (2000 + 2×40 – 2). While adding the second dimension to the MIRT model, there are 2038 (2000 + 40 – 2) more parameters to be estimated for the same data set. It is uncertain how the R^2 analog of the MIRT model reacts to the huge number of the degrees of freedom. It is also unclear how the *RLR* index reflects the potential

inflation problem for the R^2 analogs for two successive models. Even though the current findings are positive, the succeeding research should focus on the degrees of freedom of the *RLR* index to examine the possible inflation problem.

Third, simulation studies offer a means to verify the theoretical statistical properties in practice, but the simulation scenarios always have less than real complexity. It is critical to point out that all the simulated data sets in this research were based on the simple structure and they only represented the simplest cases. Future studies should also employ mixed structure to explore the statistical characteristics of the *RLR* index in correctly identifying the true dimensionality. Furthermore, the two simulation studies employed the important variables related to dimensionality. Some other potential variables, such as the effect of the guessing parameter on model-data-fit and the interaction between item-factor structure and item discrimination (the item discrimination are different for each factor) may be appealing topics for future research. Besides, the comparisons between the *RLR* index and the non-parametric indices on detecting dimensionality would be worth investigation. To detect the limitation of the *RLR* index, it would also be of interest to decide the minimum number of items and the minimum level of item discrimination representing one identifiable dimension.

Last, it is not surprising that the choice of the appropriate dimensionality assessing method is constrained by the limitations of estimation theory and the computer program (Tate, 2002). When using full-information factor analysis (TESTFACT), the number of factors should not exceed five in order to ensure the accuracy of the results (Bock et al., 1988). In order to demonstrate how the *RLR* index functions for under-fit, good fit, and over-fit, the maximum number of data dimensionality simulated in this research is three.

It is expected that the investigation of higher-dimensional data may be possible when a more powerful mathematical algorithm or a computer program is developed. Hopefully, the results presented in this research will offer useful information to practitioners interested in using the MIRT model. It is hoped that these findings will promote future research in this area and lead to helpful guidelines with respect to the assessment of the data dimensionality.

APPENDIX A

Mathematical Derivation of Esrella's (1998) R^2 Analog

solve $\dfrac{d\phi}{1-\phi} = \dfrac{dA}{1-\dfrac{A}{B}}$

Solution:

$$\frac{d\phi}{1-\phi} = \frac{dA}{1-\frac{A}{B}}$$

$$\Rightarrow \int \frac{1}{1-\phi} d\phi = \int \frac{1}{(1-\frac{A}{B})} dA$$

$$\Rightarrow -\ln(1-\phi) = -B\ln(1-\frac{A}{B}) + C \text{, where C is a constant}$$

$$\Rightarrow \ln(1-\phi) = \ln(1-\frac{A}{B})^B - C$$

$$\Rightarrow (1-\phi) = (1-\frac{A}{B})^B \times \exp(-C)$$

Given that $\phi_0(0) = 0$, which means when A=0, ϕ =0

$$\Rightarrow 1-0 = (1-0)^B \exp(-C)$$

$$\Rightarrow \exp(-C) = 1$$

$$\Rightarrow C = 0$$

Thus $1-\phi = (1-\frac{A}{B})^B$

$$\Rightarrow \phi = 1-(1-\frac{A}{B})^B$$

APPENDIX B

The Conditional Distributions of the *RLR* Values in Simulation Study I

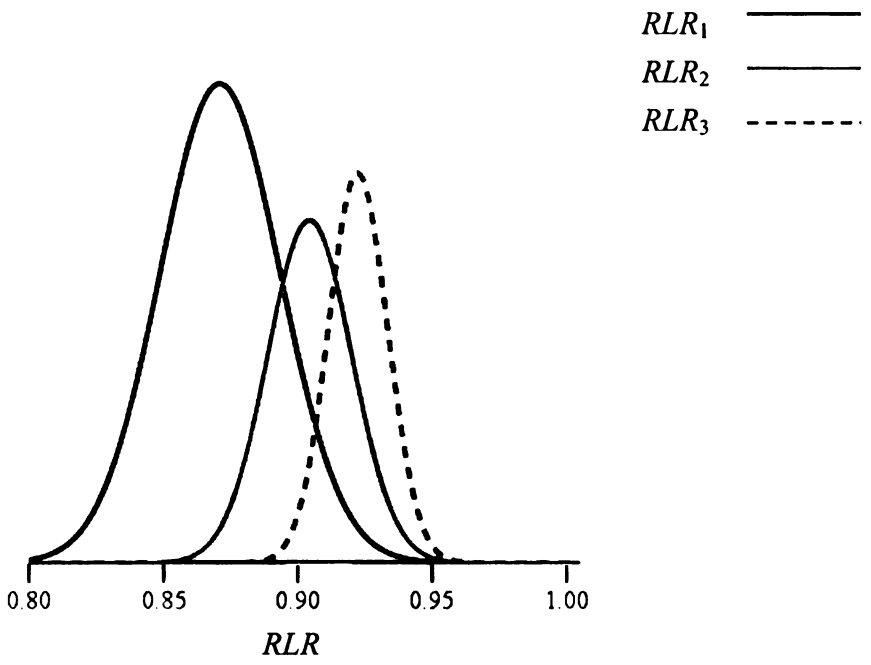


Figure B.1. The conditional distributions of the *RLR* values for Test 111with 2000 examinees

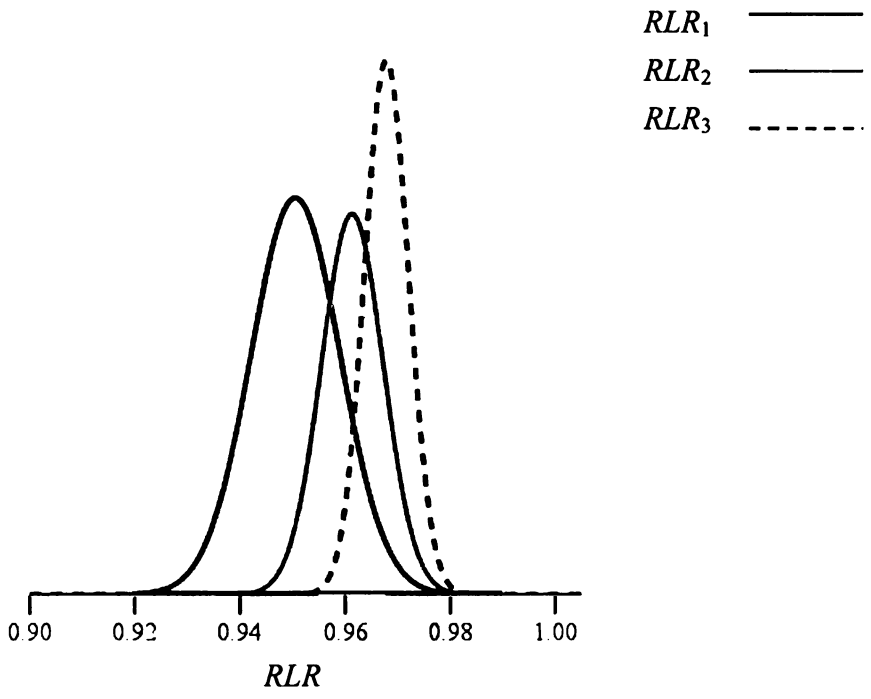


Figure B.2. The conditional distributions of the *RLR* values for Test 111with 6000 examinees

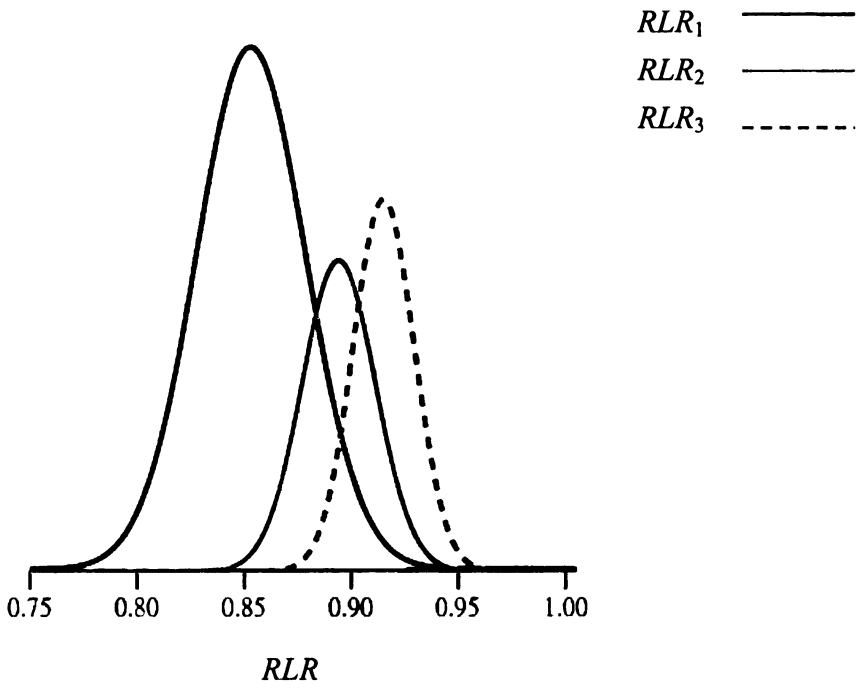


Figure B.3. The conditional distributions of the *RLR* values for Test 121with 2000 examinees

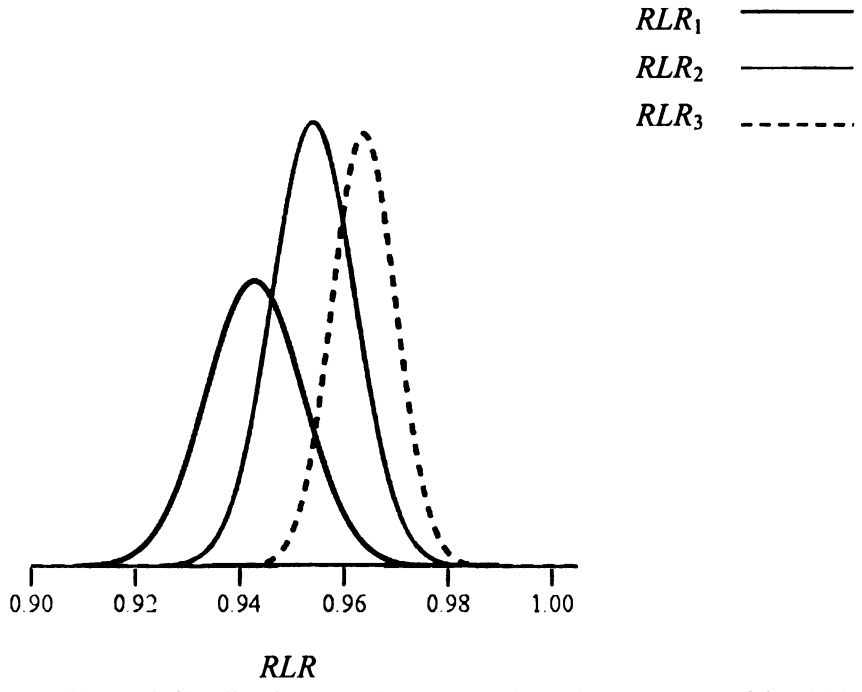


Figure B.4. The conditional distributions of the *RLR* values for Test 121with 6000 examinees

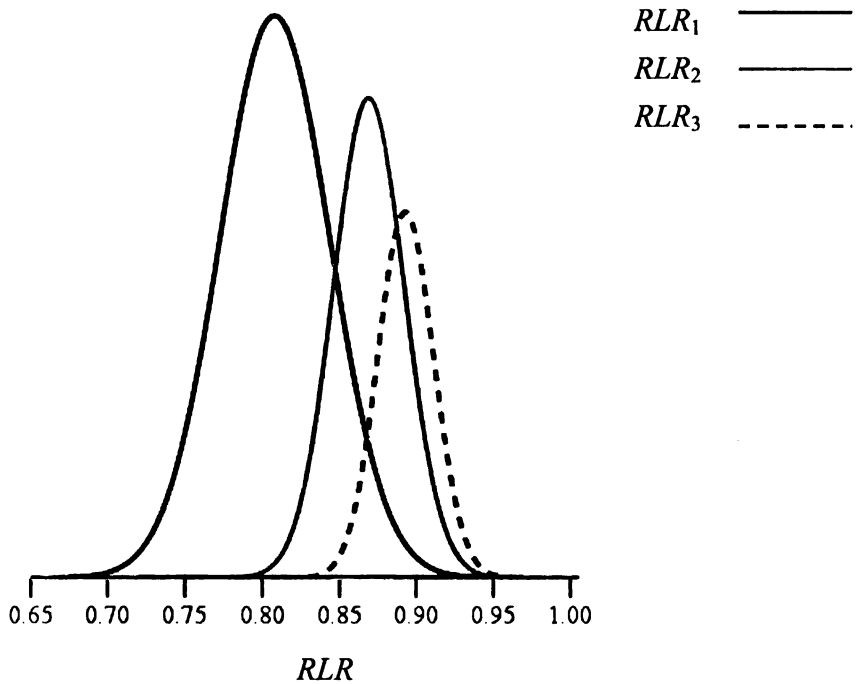


Figure B.5. The conditional distributions of the *RLR* values for Test 131with 2000 examinees

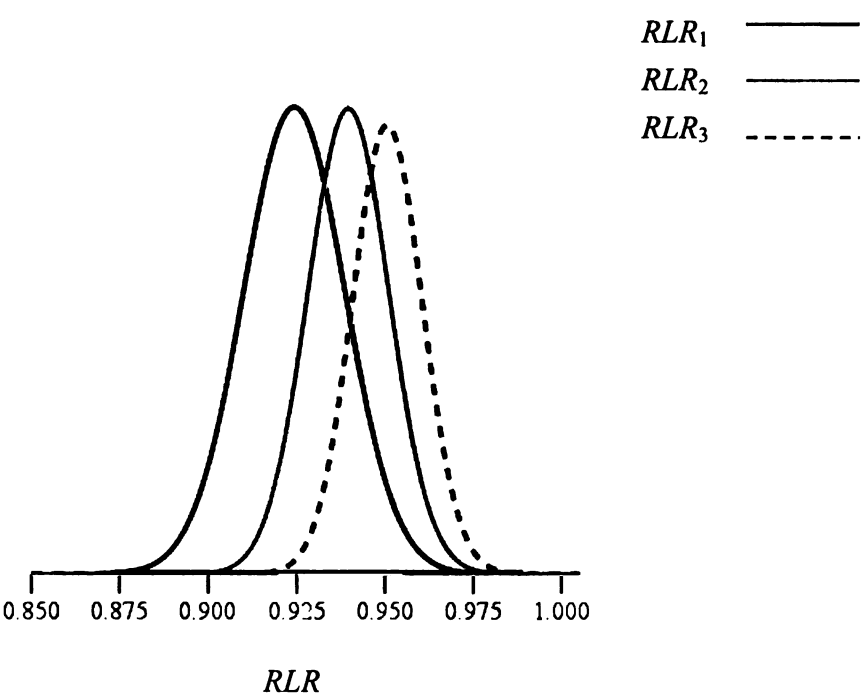


Figure B.6. The conditional distributions of the *RLR* values for Test 131with 6000 examinees

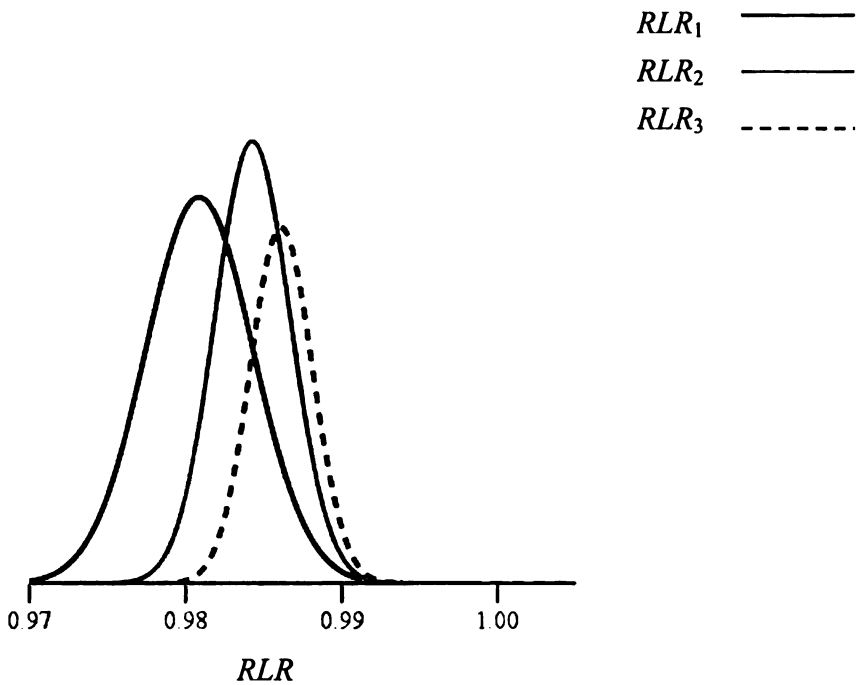


Figure B.7. The conditional distributions of the *RLR* values for Test 211with 2000 examinees

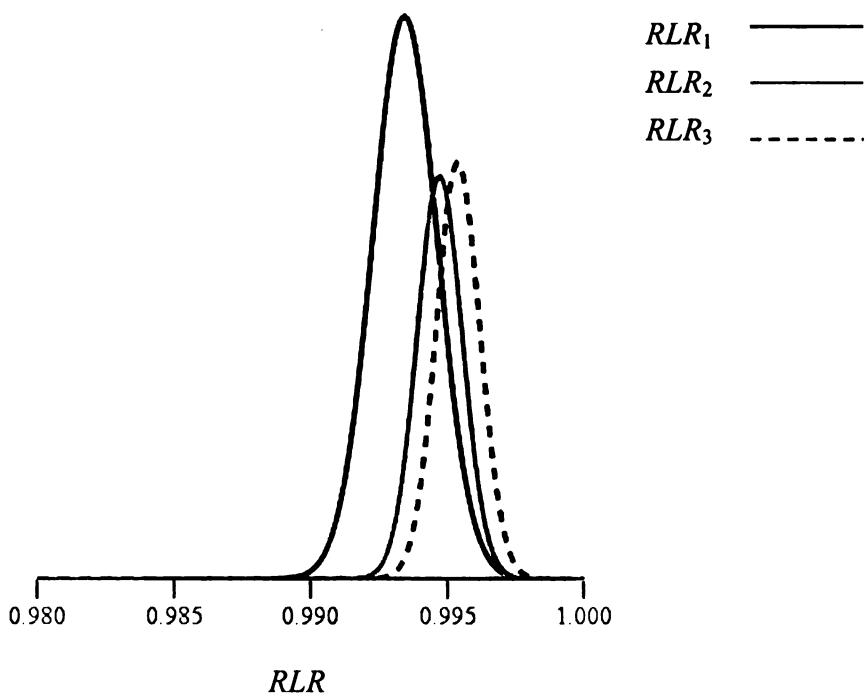


Figure B.8. The conditional distributions of the *RLR* values for Test 211with 6000 examinees

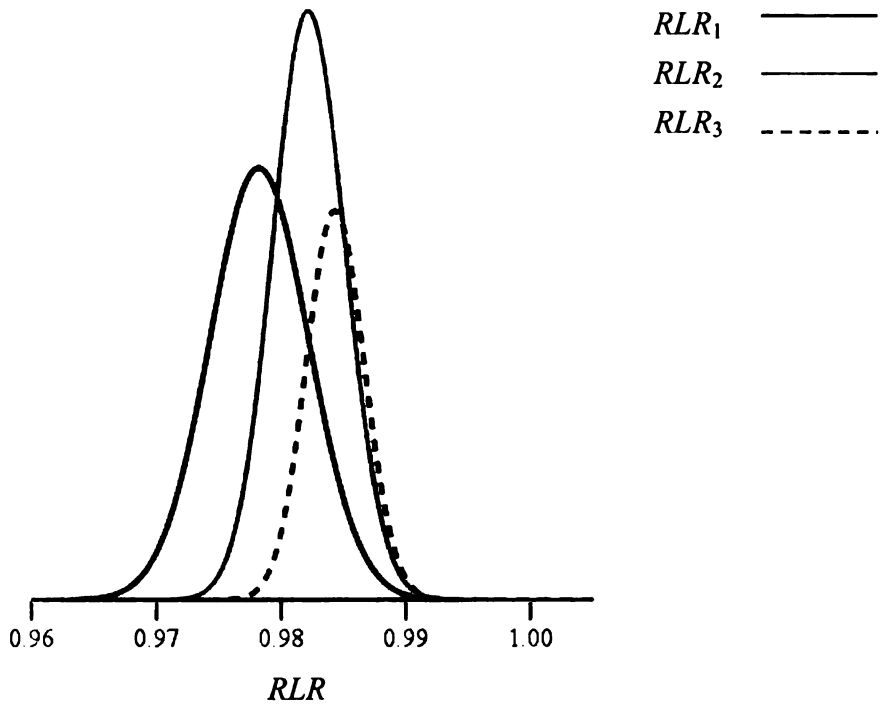


Figure B.9. The conditional distributions of the RLR values for Test 221with 2000 examinees

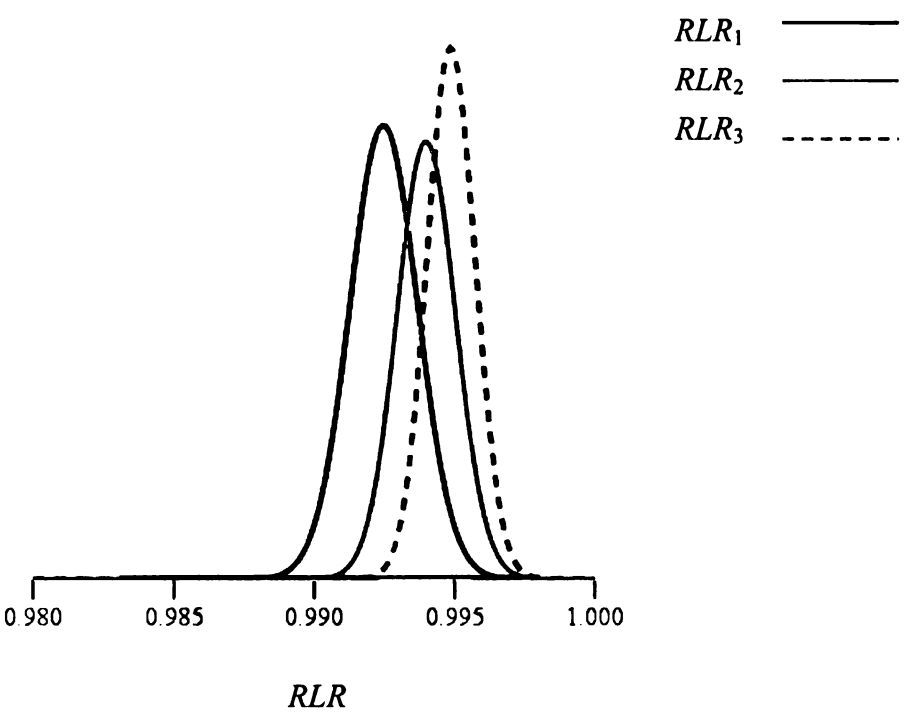


Figure B.10. The conditional distributions of the *RLR* values for Test 221with 6000 examinees

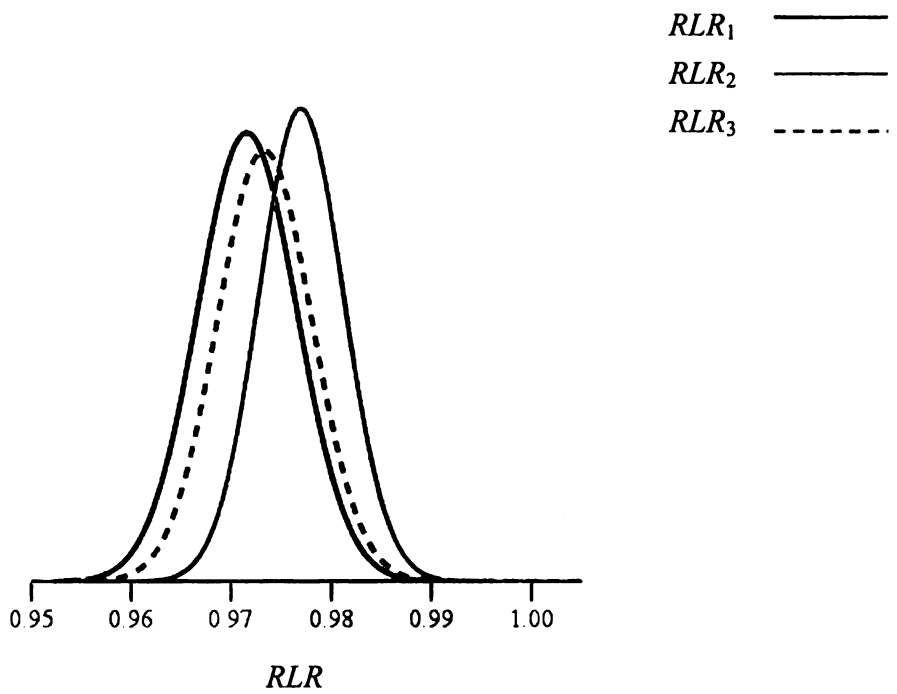


Figure B.11. The conditional distributions of the *RLR* values for Test 231with 2000 examinees

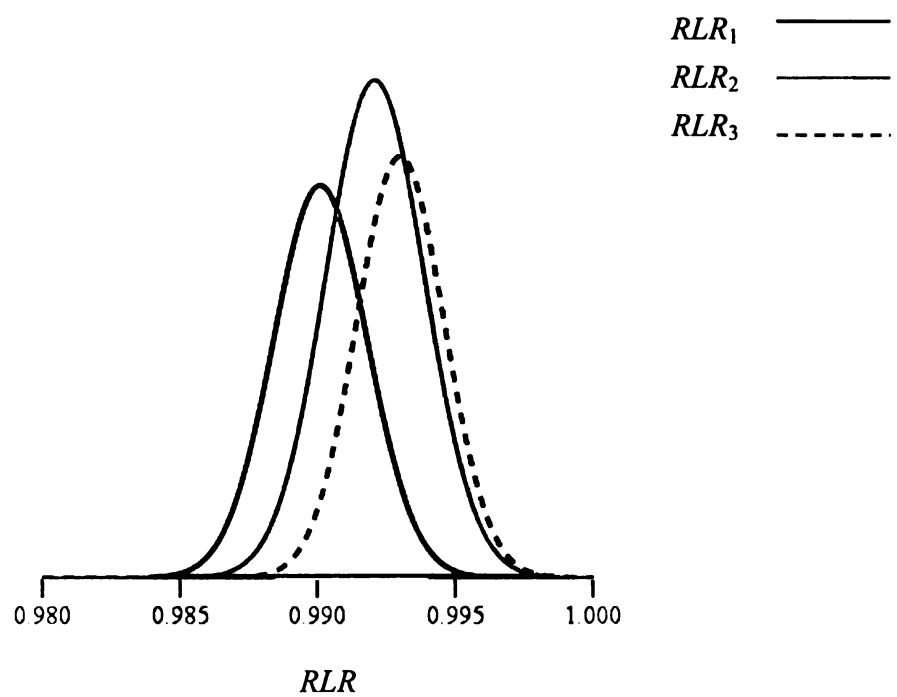


Figure B.12. The conditional distributions of the *RLR* values for Test 231with 6000 examinees

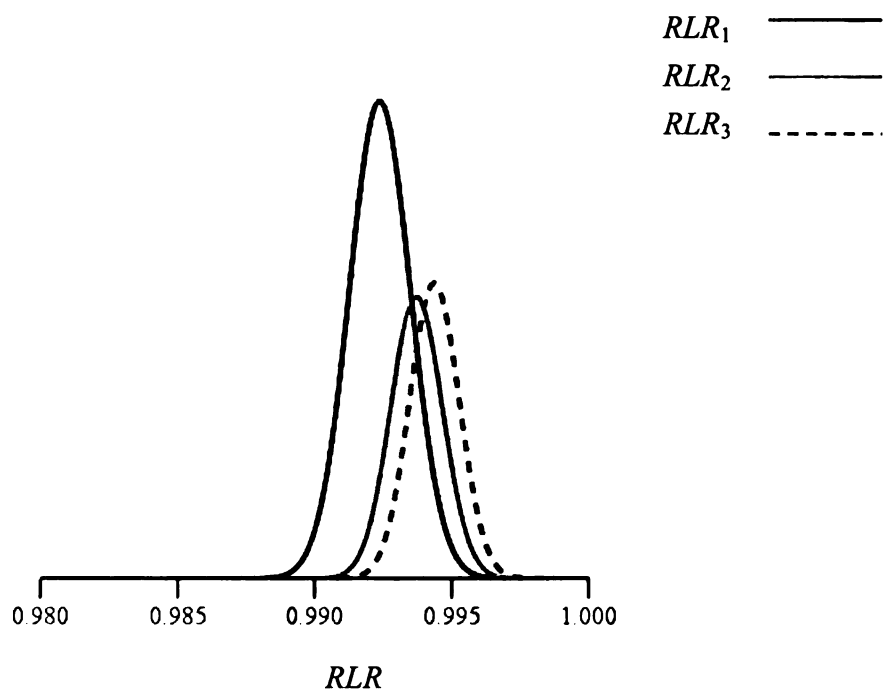


Figure B.13. The conditional distributions of the *RLR* values for Test 331with 2000 examinees

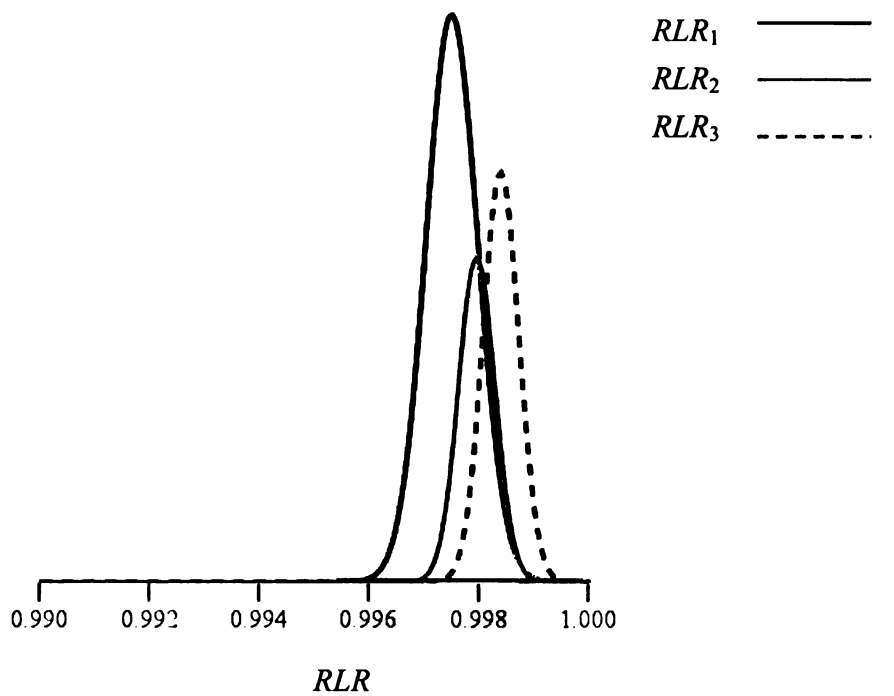


Figure B.14. The conditional distributions of the *RLR* values for Test 331with 6000 examinees

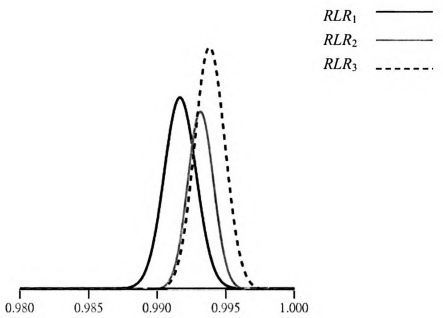


Figure B.15. The conditional distributions of the *RLR* values for Test 321with 2000 examinees

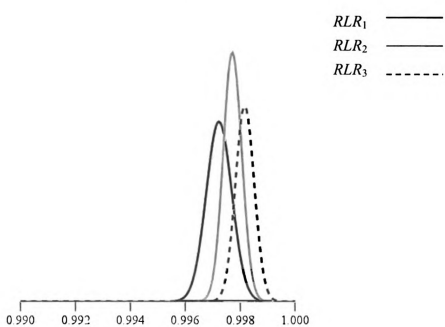


Figure B.16. The conditional distributions of the *RLR* values for Test 321with 6000 examinees

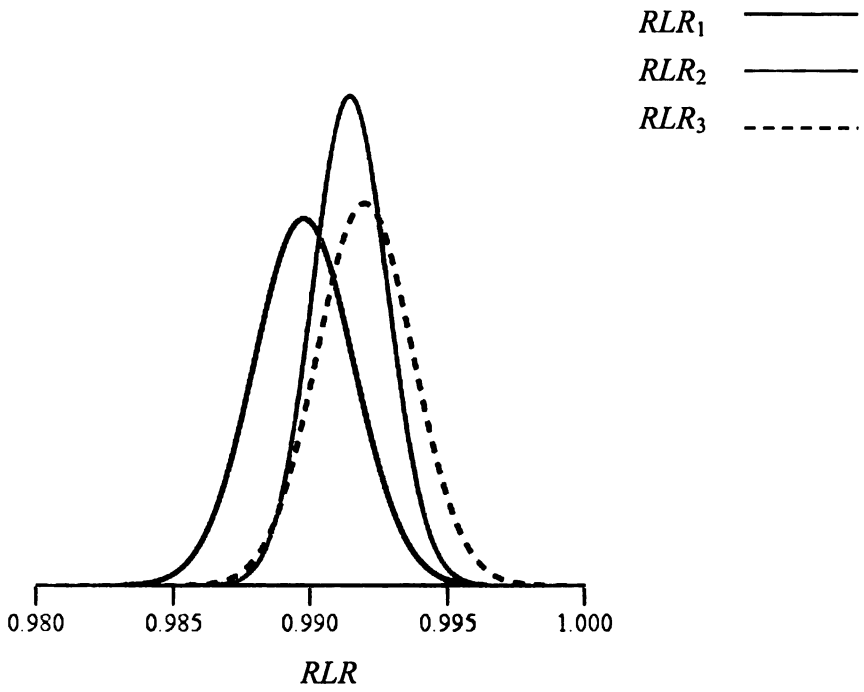


Figure B.17. The conditional distributions of the *RLR* values for Test 331with 2000 examinees

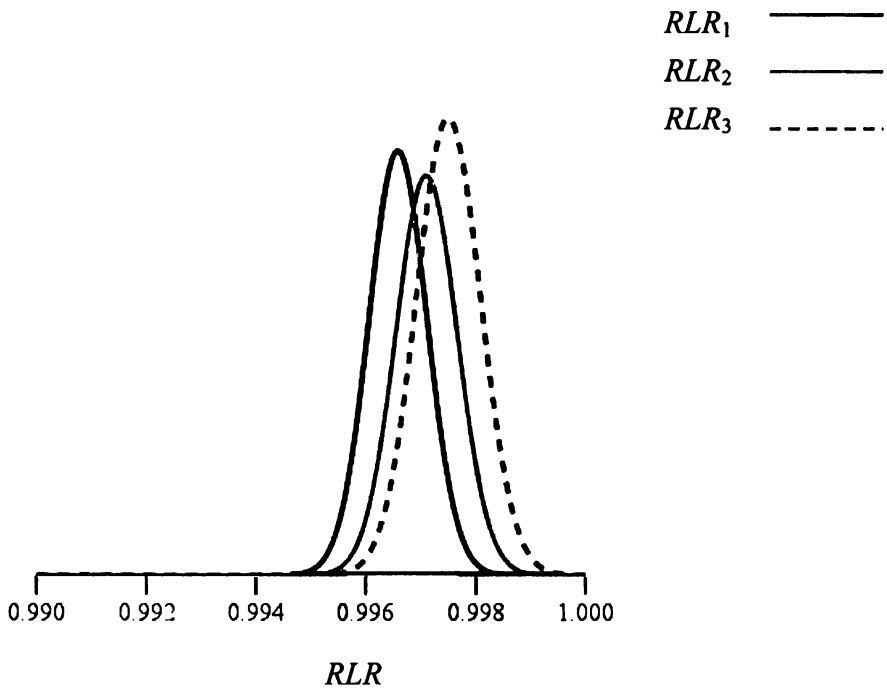


Figure B.18. The conditional distributions of the *RLR* values for Test 331with 6000 examinees

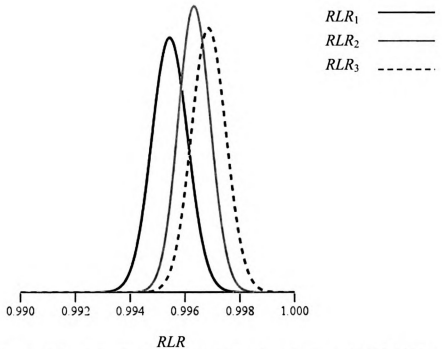


Figure B.19. The conditional distributions of the *RLR* values for Test 411with 2000 examinees

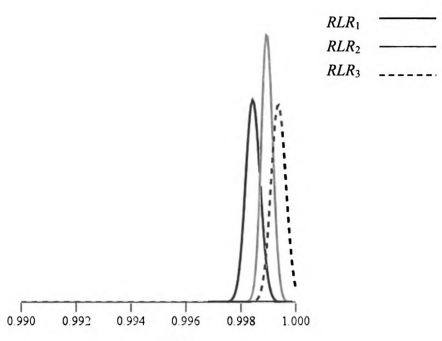


Figure B.20. The conditional distributions of the *RLR* values for Test 411with 6000 examinees

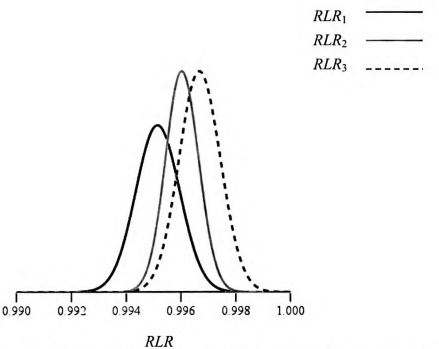


Figure B.21. The conditional distributions of the *RLR* values for Test 421with 2000 examinees

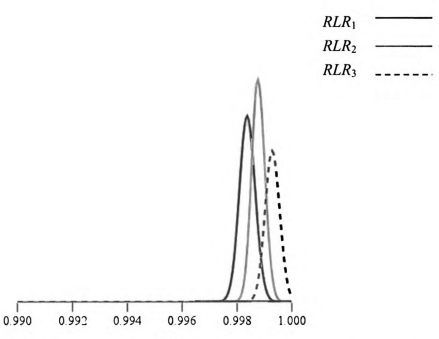


Figure B.22. The conditional distributions of the *RLR* values for Test 421with 6000 examinees

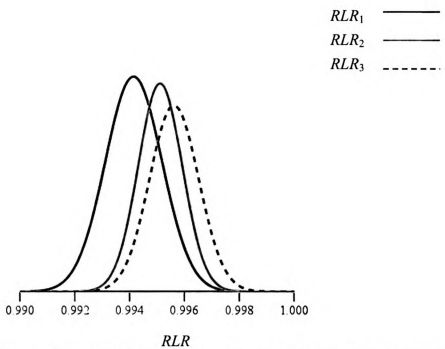


Figure B.23. The conditional distributions of the *RLR* values for Test 431with 2000 examinees

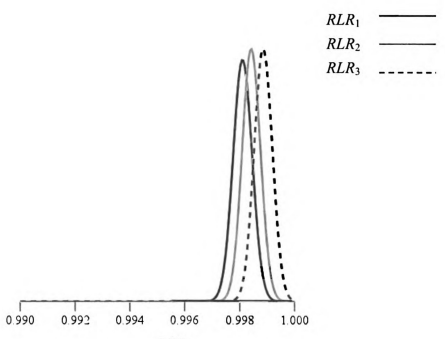


Figure B.24. The conditional distributions of the *RLR* values for Test 431with 6000 examinees

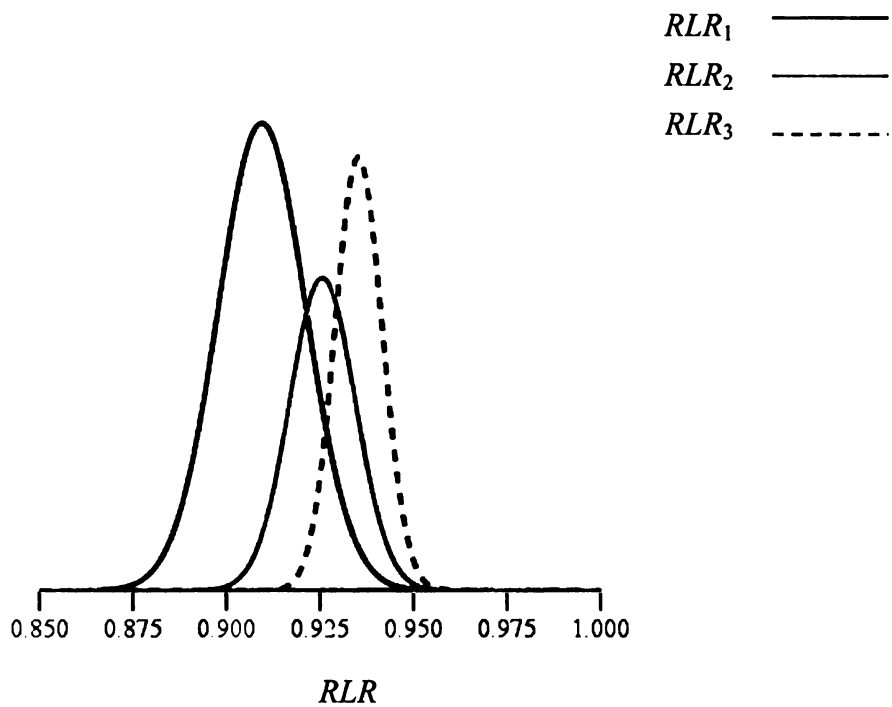


Figure B.25. The conditional distributions of the *RLR* values for Test 112 with 2000 examinees

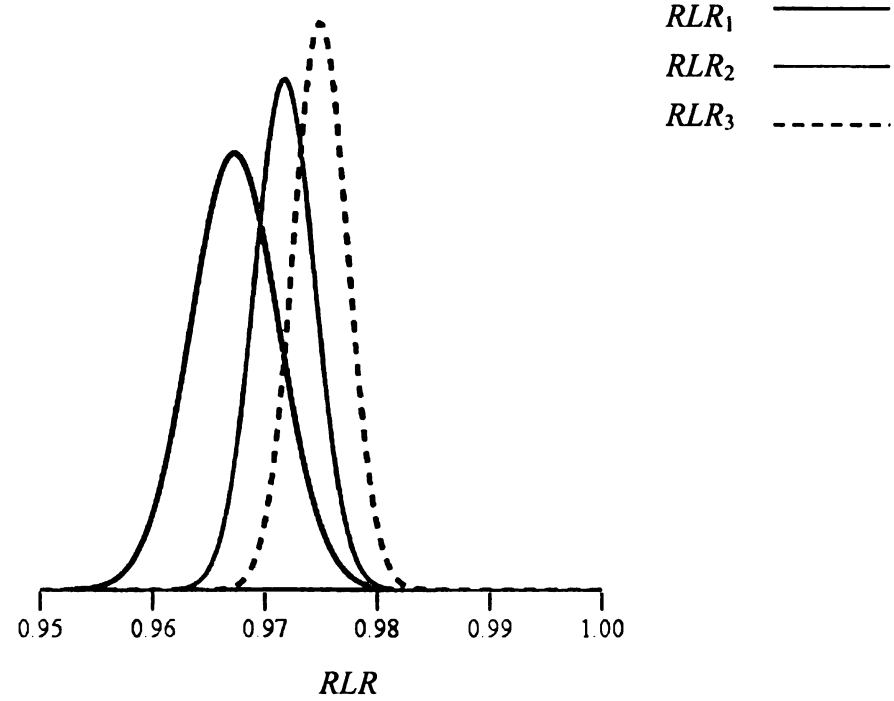


Figure B.26. The conditional distributions of the *RLR* values for Test 112 with 6000 examinees

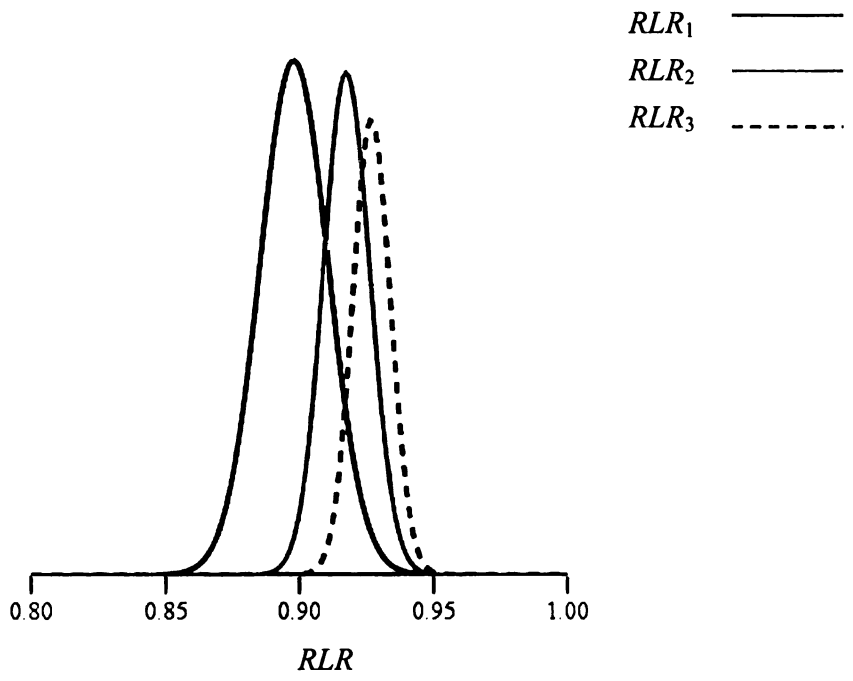


Figure B.27. The conditional distributions of the *RLR* values for Test 122 with 2000 examinees

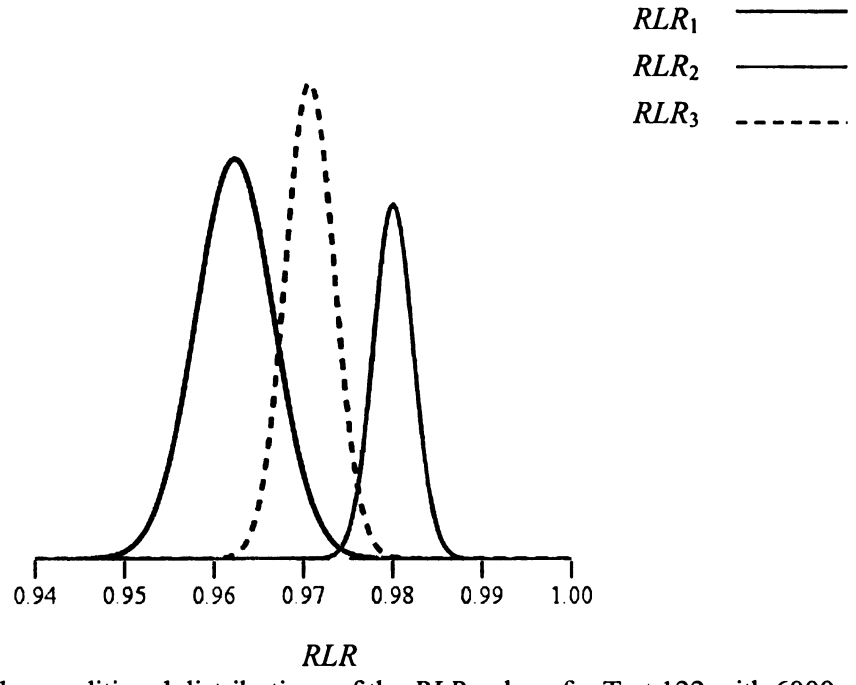


Figure B.28. The conditional distributions of the *RLR* values for Test 122 with 6000 examinees

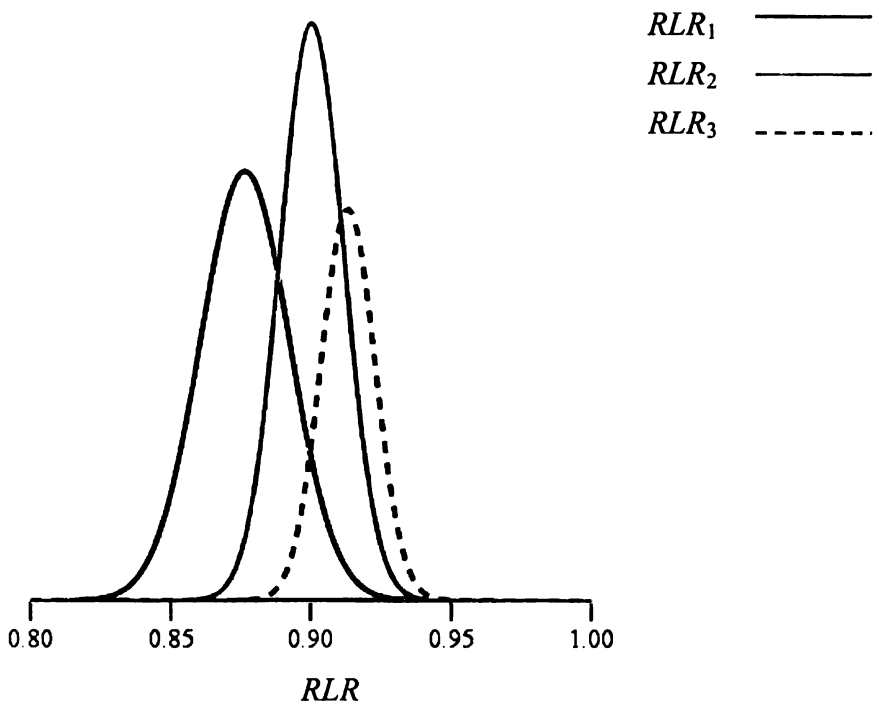


Figure B.29. The conditional distributions of the *RLR* values for Test 132 with 2000 examinees

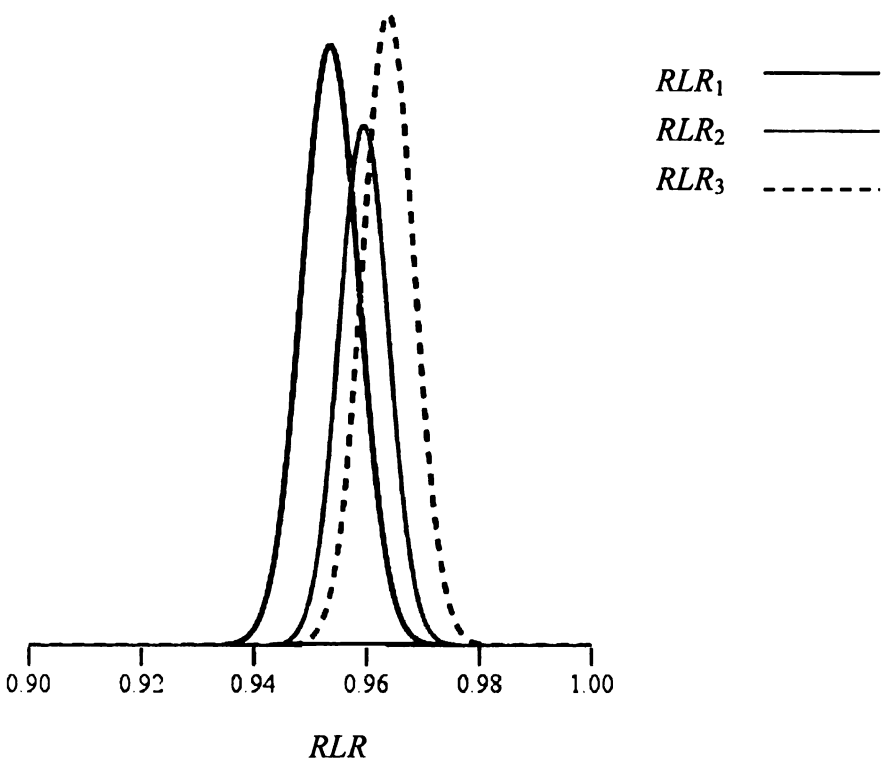


Figure B.30. The conditional distributions of the *RLR* values for Test 132 with 6000 examinees

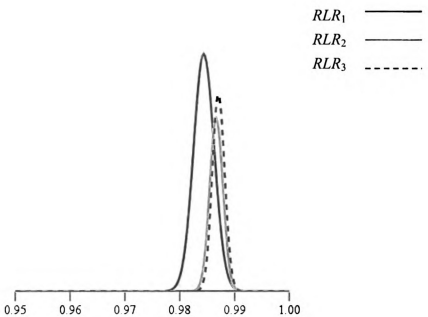


Figure B.31. The conditional distributions of the RLR values for Test 212 with 2000 examinees

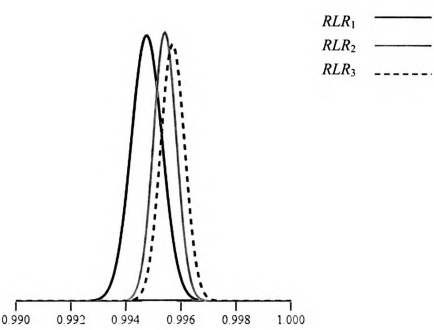


Figure B.32. The conditional distributions of the *RLR* values for Test 212 with 6000 examinees

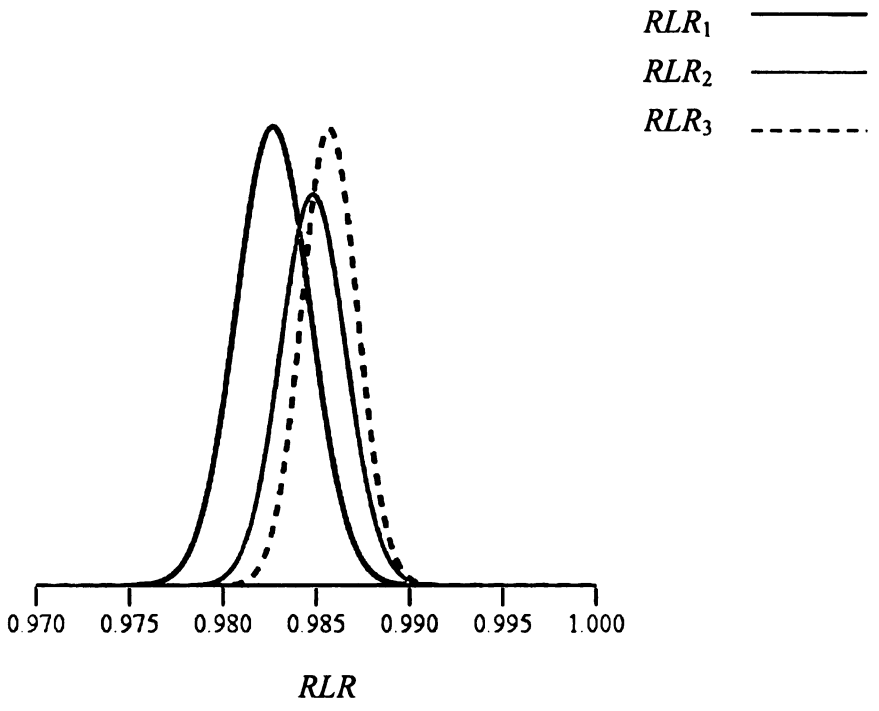


Figure B.33. The conditional distributions of the *RLR* values for Test 222 with 2000 examinees

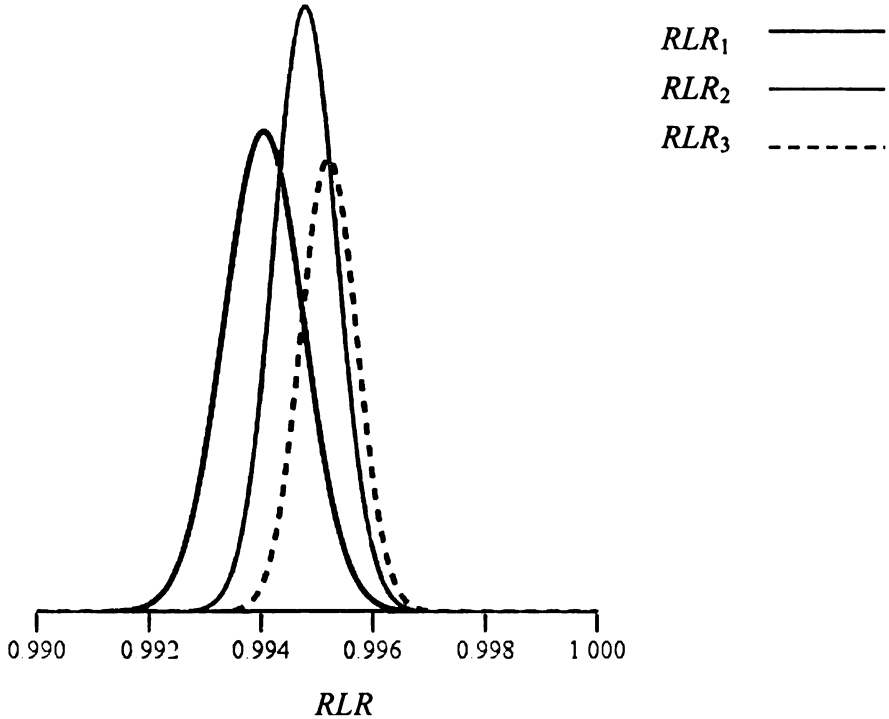


Figure B.34. The conditional distributions of the *RLR* values for Test 222 with 6000 examinees

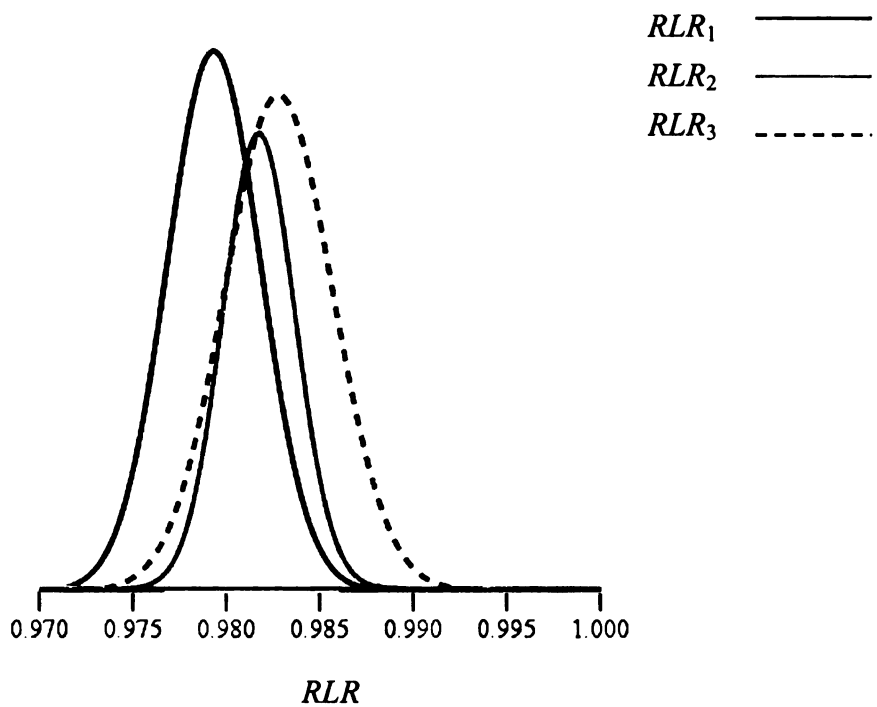


Figure B.35. The conditional distributions of the *RLR* values for Test 232 with 2000 examinees

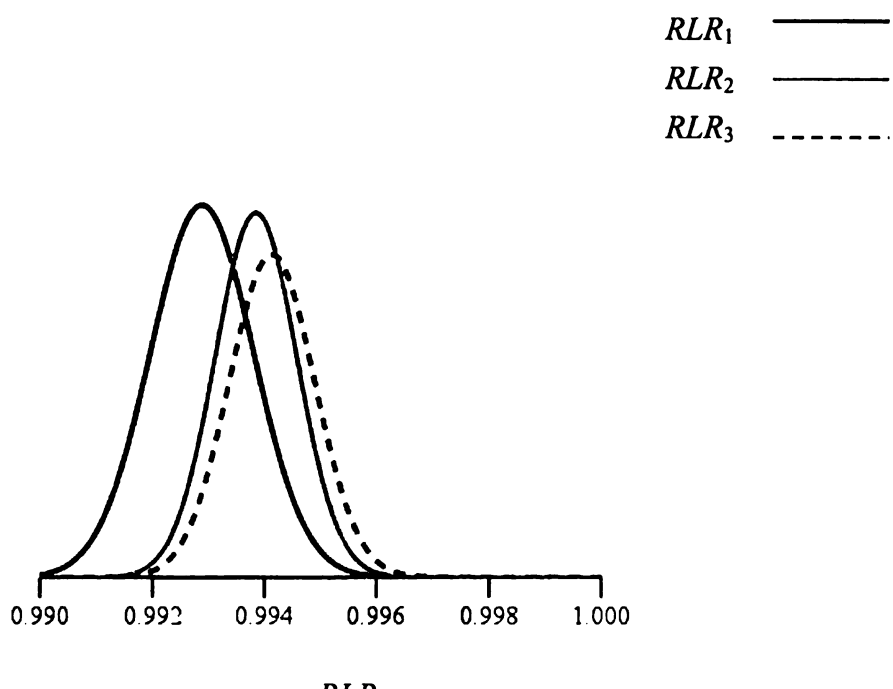


Figure B.36. The conditional distributions of the *RLR* values for Test 232 with 6000 examinees

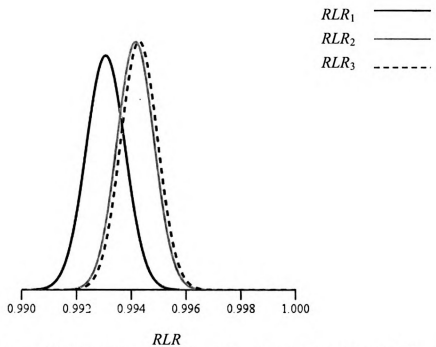


Figure B.37: The conditional distributions of the *RLR* values for Test 312 with 2000 examinees

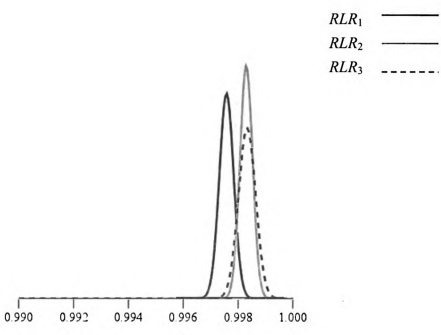


Figure B.38: The conditional distributions of the *RLR* values for Test 312 with 6000 examinees

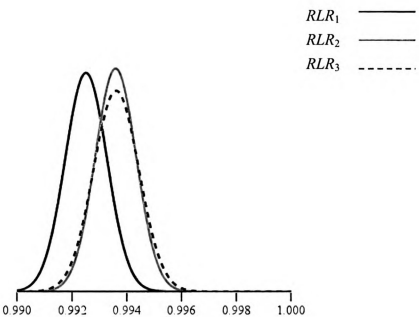


Figure B.39. The conditional distributions of the *RLR* values for Test 322 with 2000 examinees

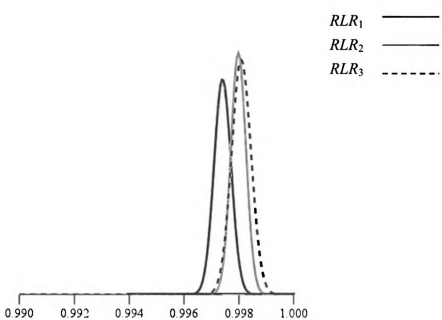


Figure B.40. The conditional distributions of the *RLR* values for Test 322 with 6000 examinees

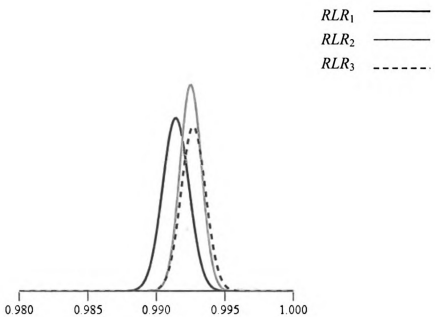


Figure B.41. The conditional distributions of the *RLR* values for Test 332 with 2000 examinees

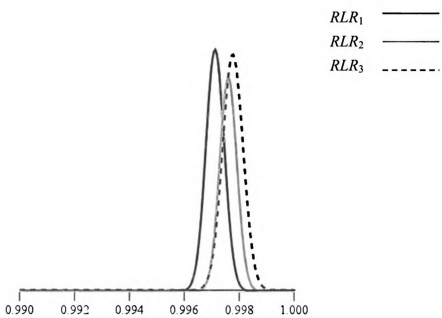


Figure B.42. The conditional distributions of the *RLR* values for Test 332 with 6000 examinees

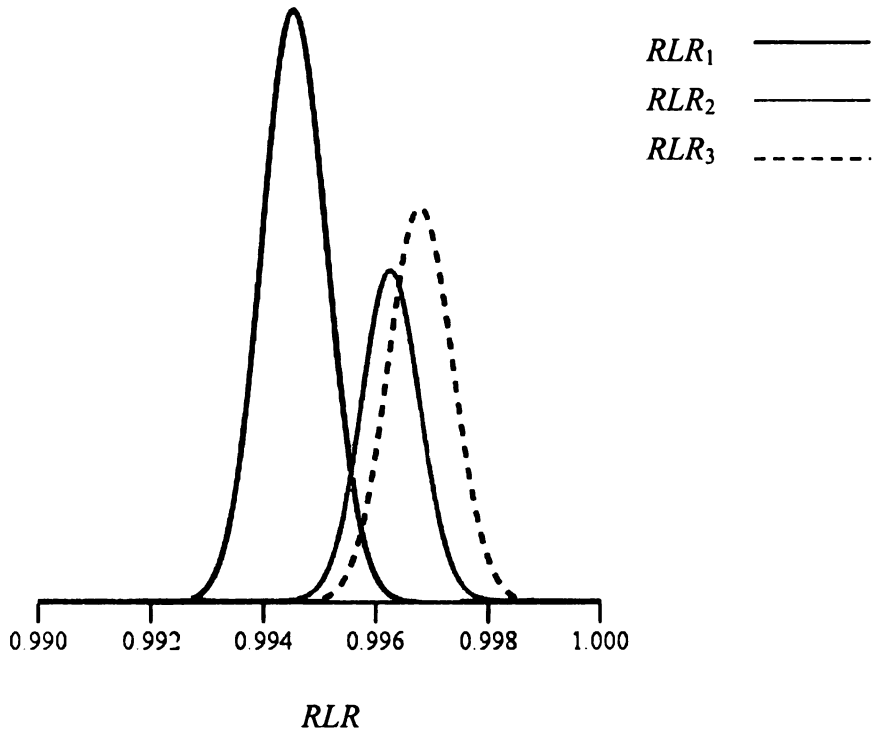


Figure B.43. The conditional distributions of the *RLR* values for Test 412 with 2000 examinees

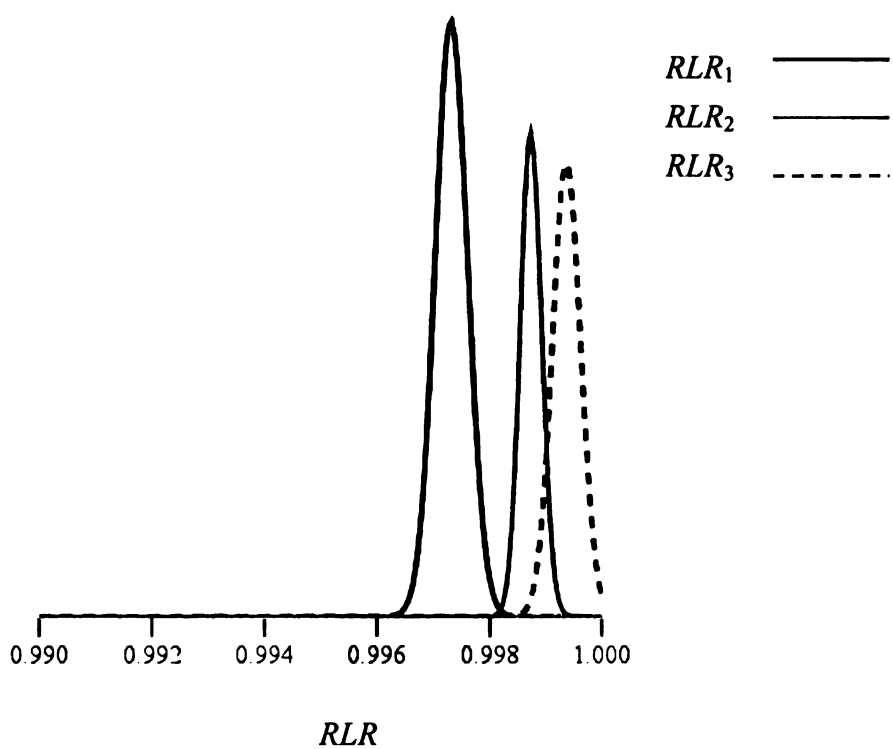


Figure B.44. The conditional distributions of the *RLR* values for Test 412 with 6000 examinees

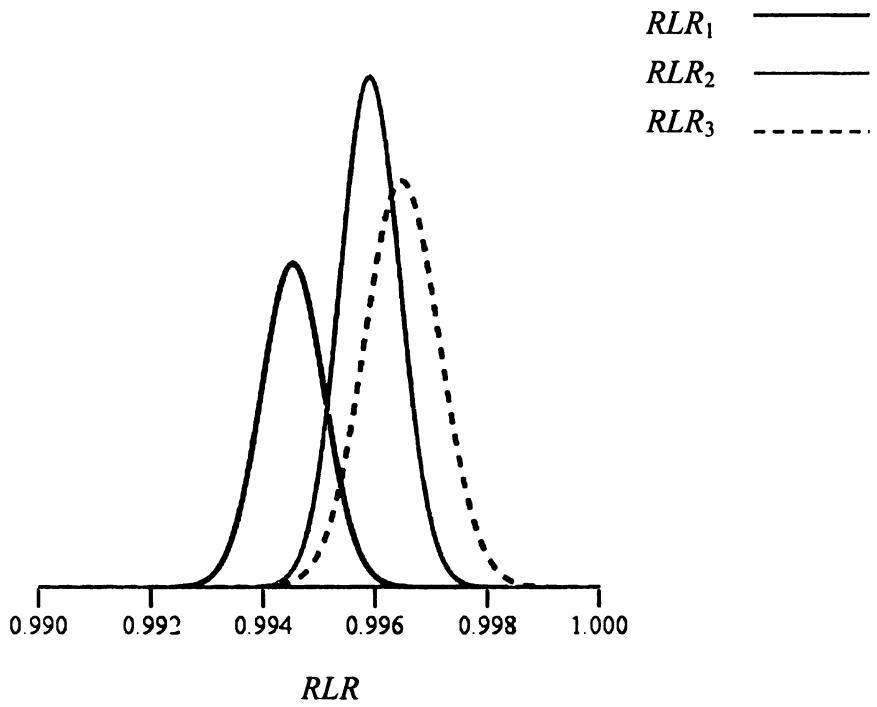


Figure B.45. The conditional distributions of the *RLR* values for Test 422 with 2000 examinees

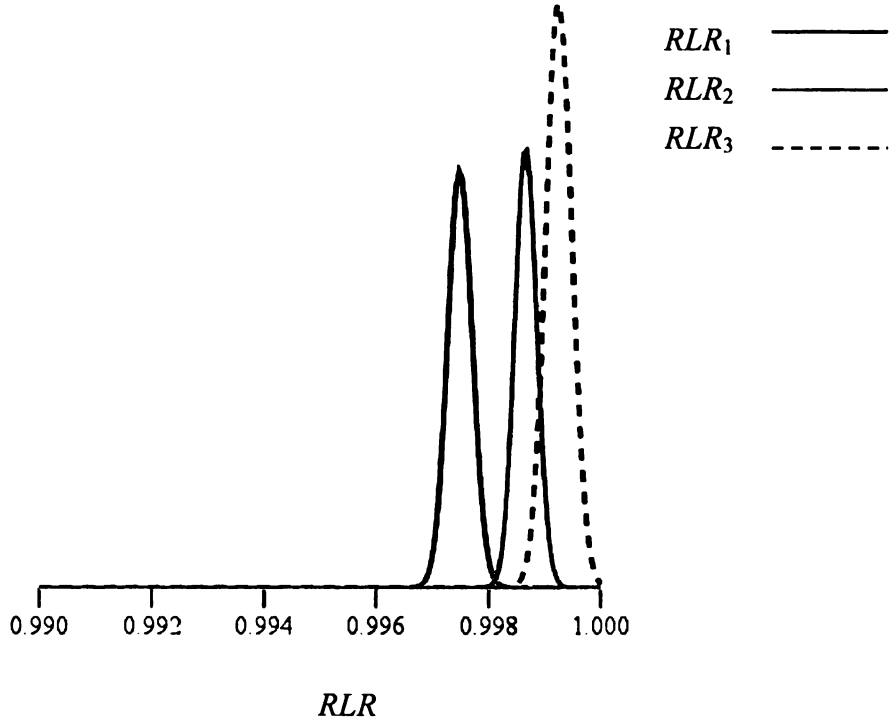


Figure B.46. The conditional distributions of the *RLR* values for Test 422 with 6000 examinees

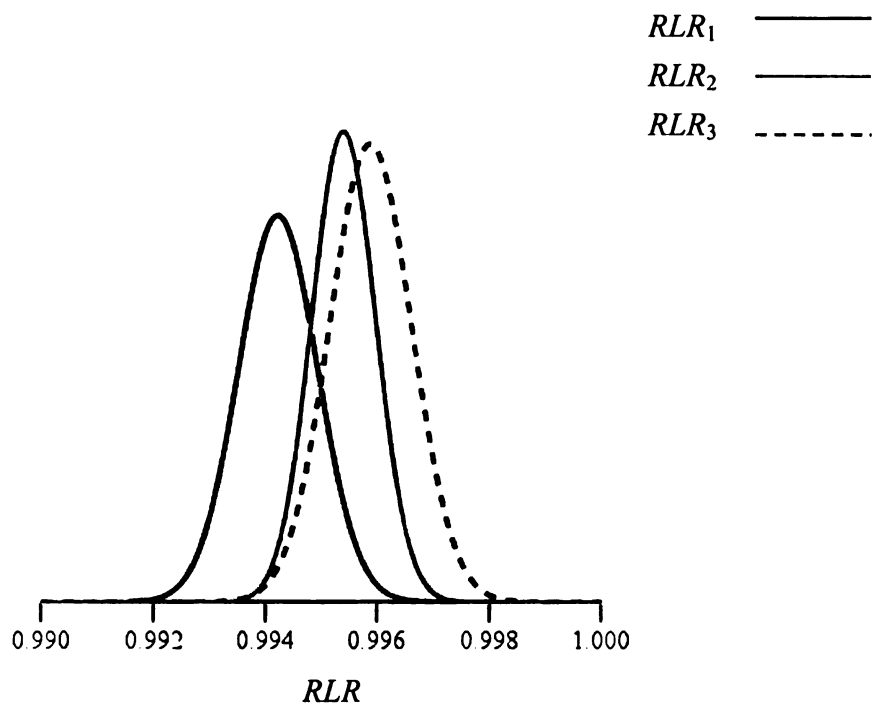


Figure B.47. The conditional distributions of the *RLR* values for Test 432 with 2000 examinees

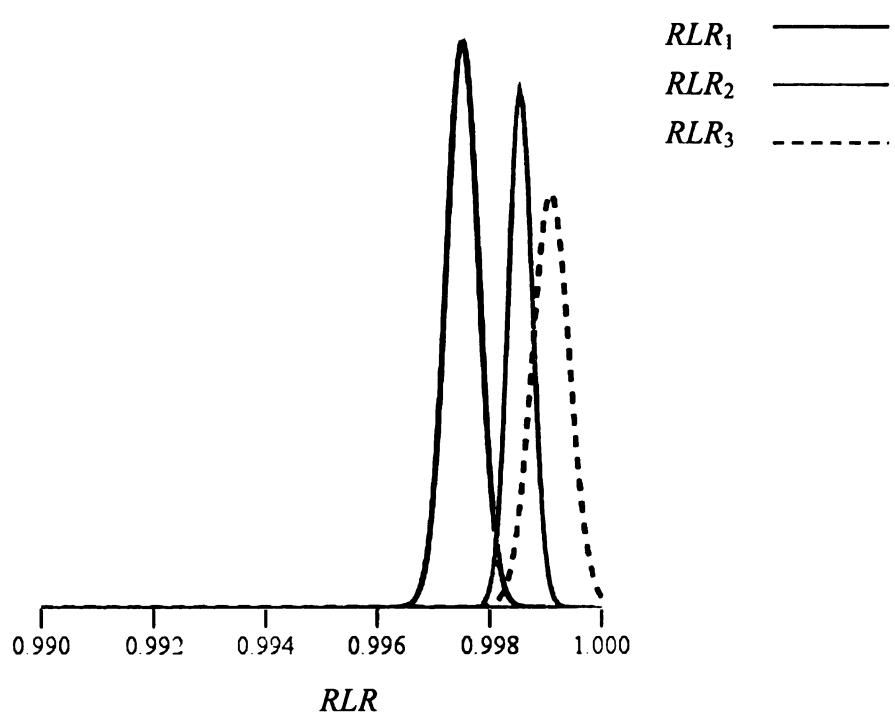


Figure B.48. The conditional distributions of the *RLR* values for Test 432 with 6000 examinees

APPENDIX C

The Conditional Distributions of the *RLR* Values in Simulation Study II

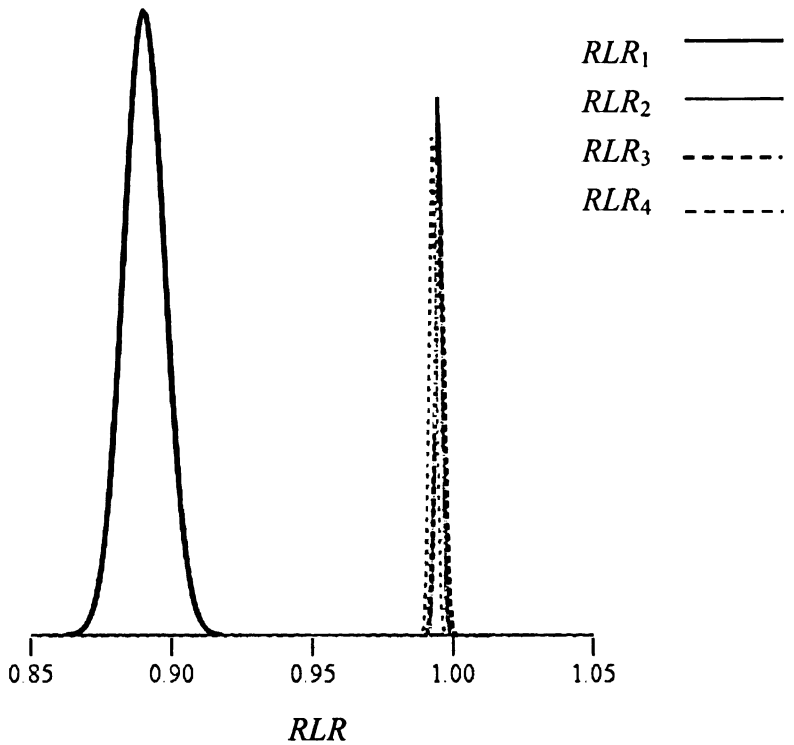


Figure C.1. The conditional distributions of the *RLR* values for Form 111

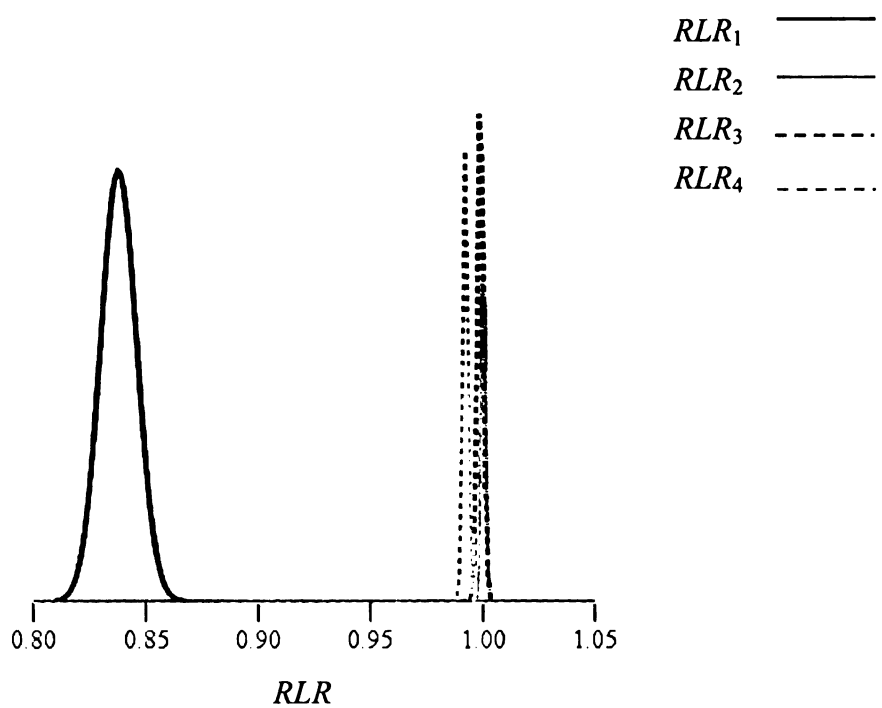


Figure C.2. The conditional distributions of the *RLR* values for Form 112

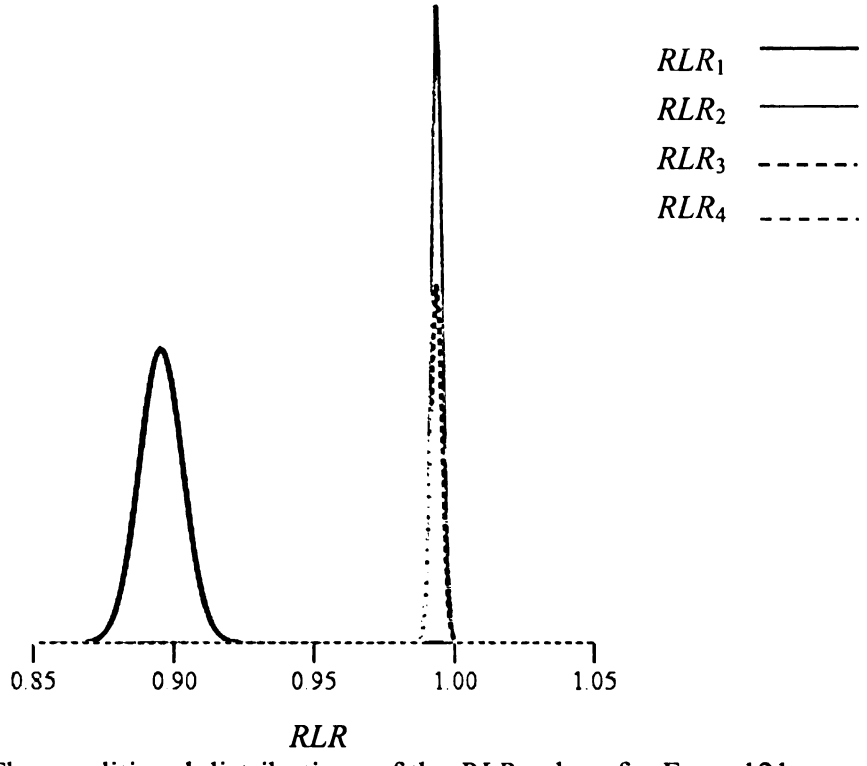


Figure C.3. The conditional distributions of the *RLR* values for Form 121

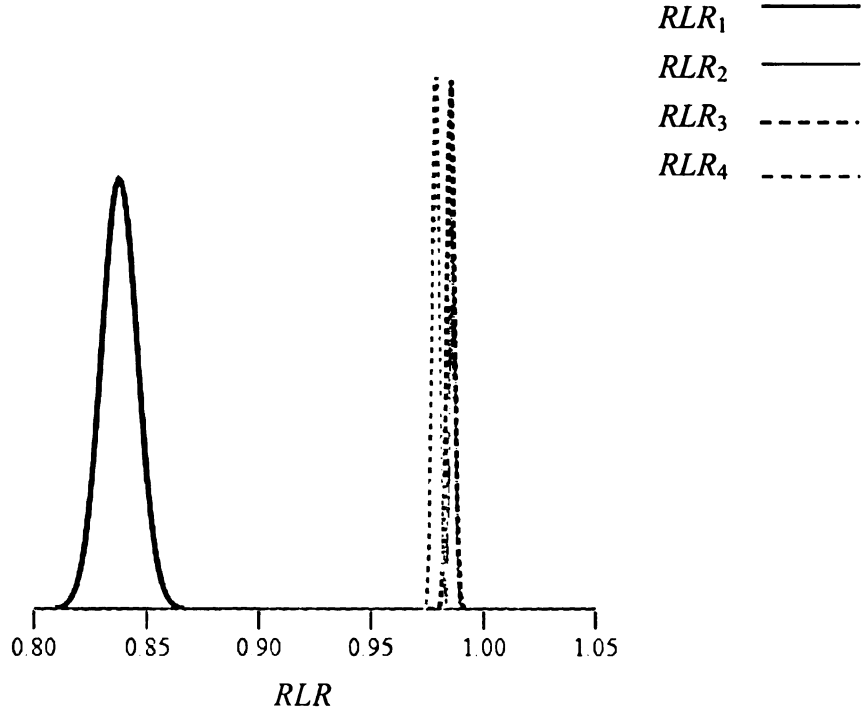


Figure C.4. The conditional distributions of the *RLR* values for Form 122

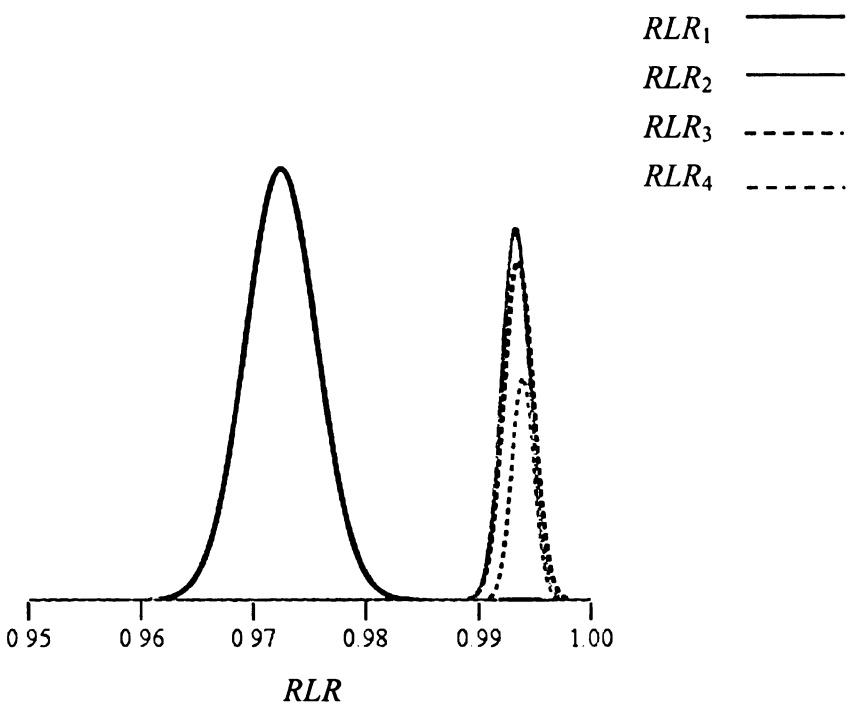


Figure C.5. The conditional distributions of the *RLR* values for Form 131

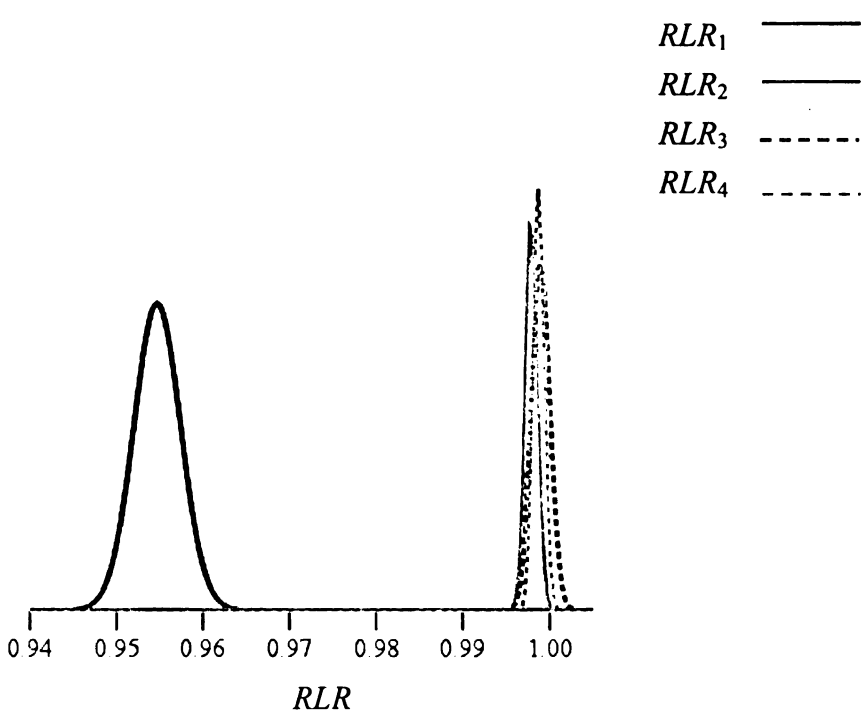


Figure C.6. The conditional distributions of the *RLR* values for Form 132

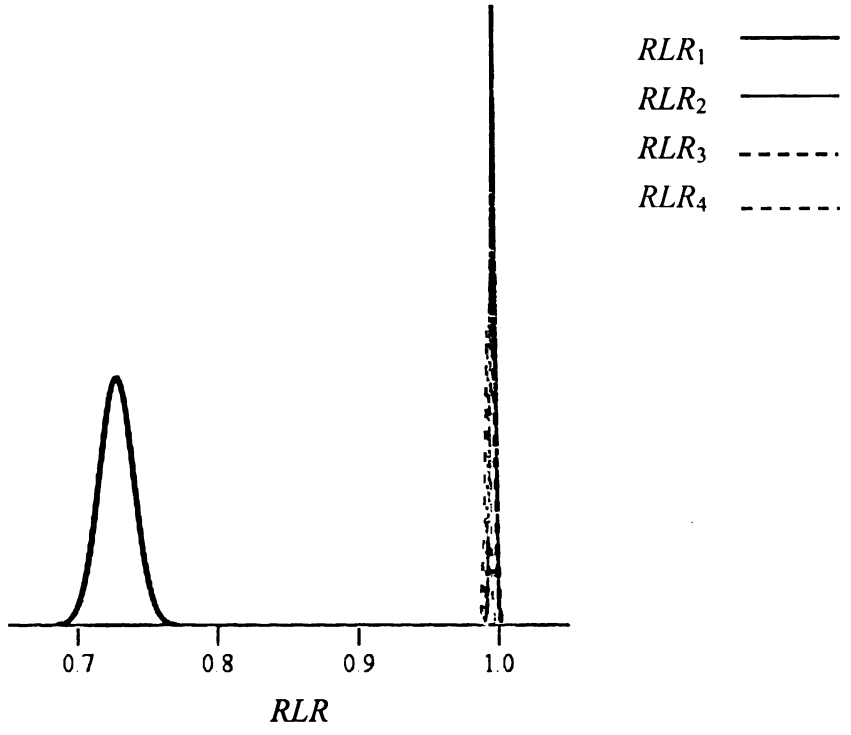


Figure C.7. The conditional distributions of the *RLR* values for Form 211

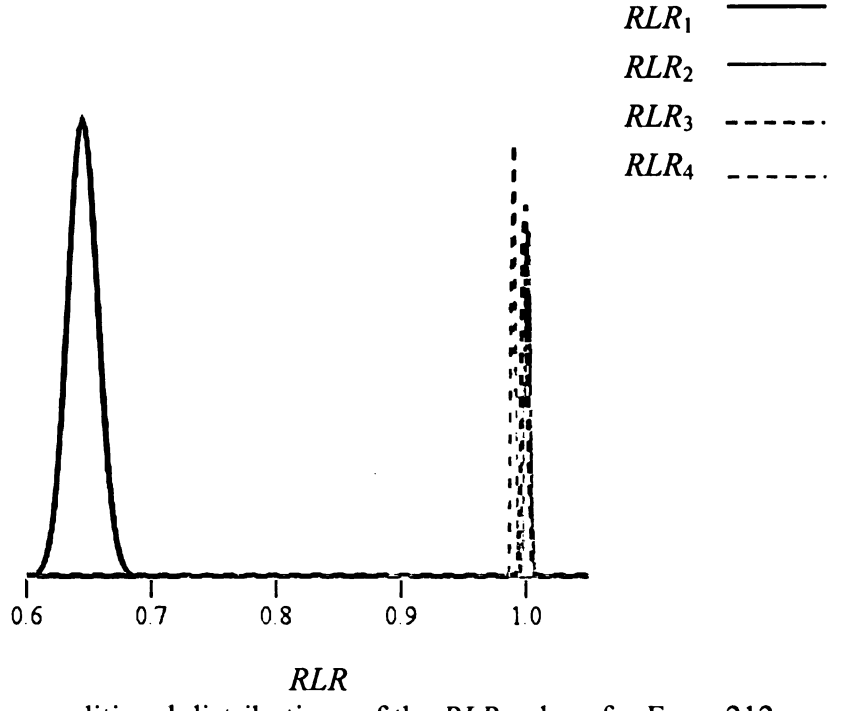


Figure C.8. The conditional distributions of the *RLR* values for Form 212

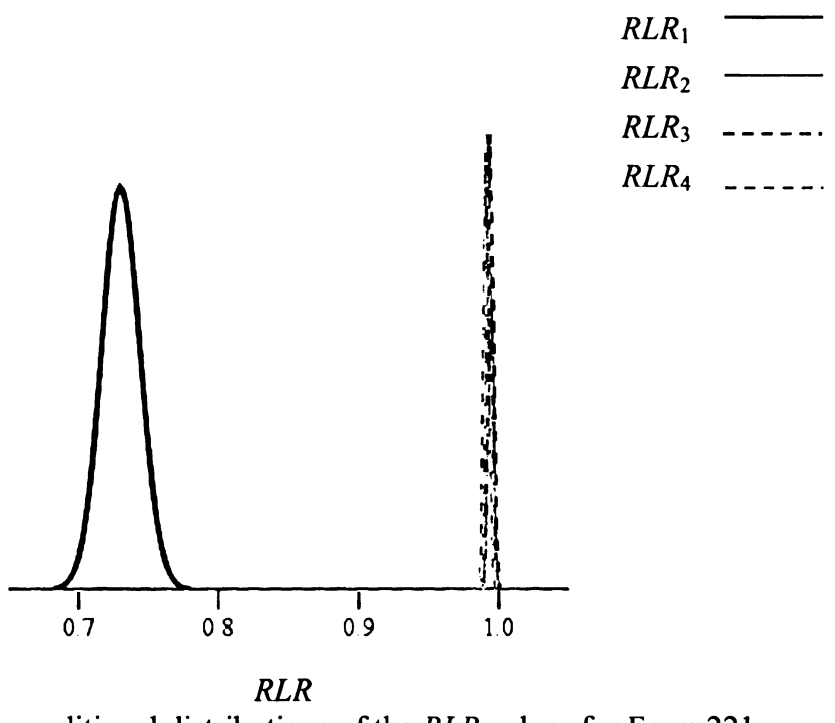


Figure C.9. The conditional distributions of the *RLR* values for Form 221

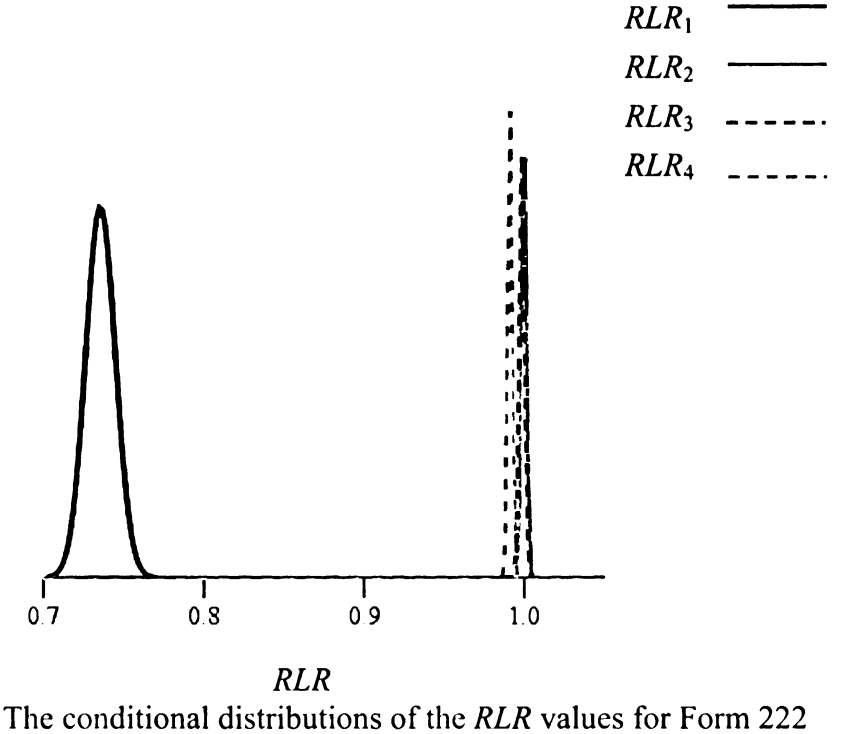


Figure C.10. The conditional distributions of the *RLR* values for Form 222

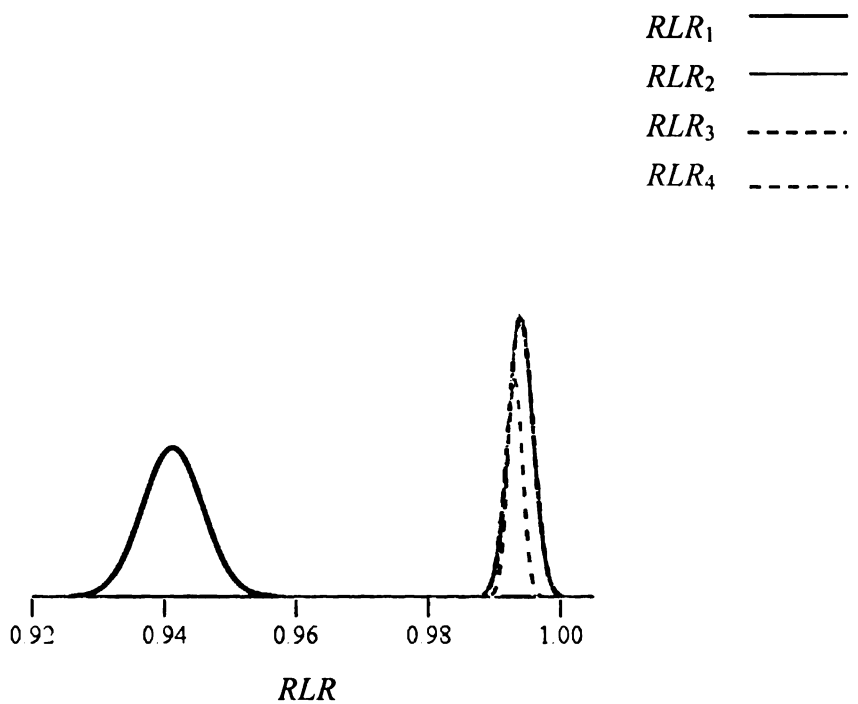


Figure C.11. The conditional distributions of the *RLR* values for Form 231

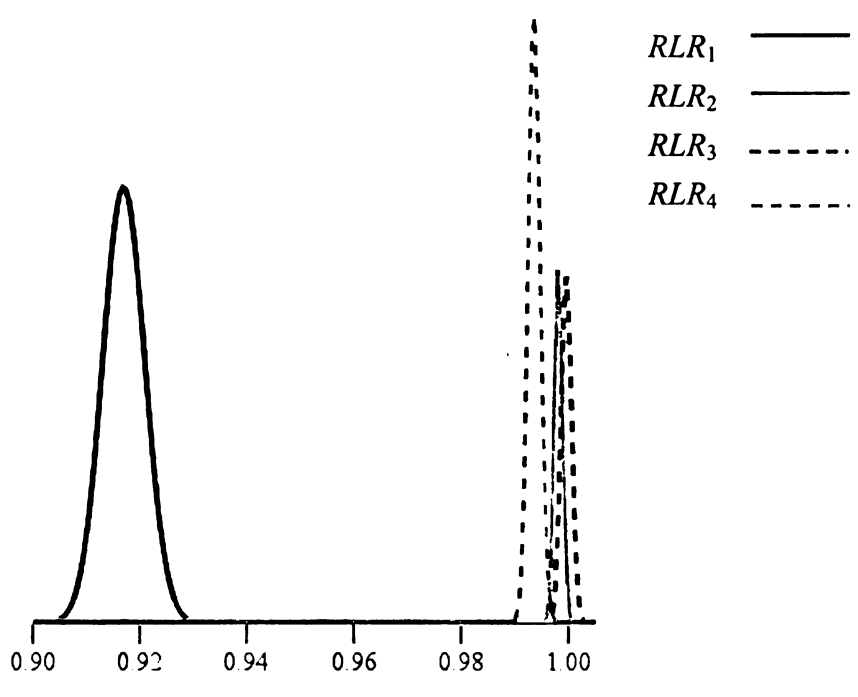


Figure C.12. The conditional distributions of the *RLR* values for Form 232

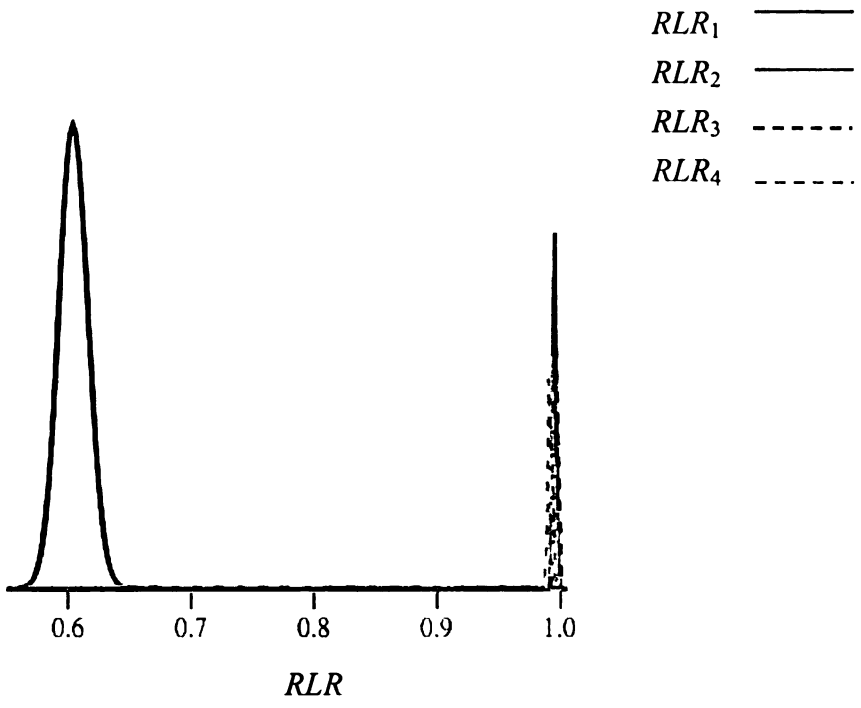


Figure C.13. The conditional distributions of the *RLR* values for Form 311

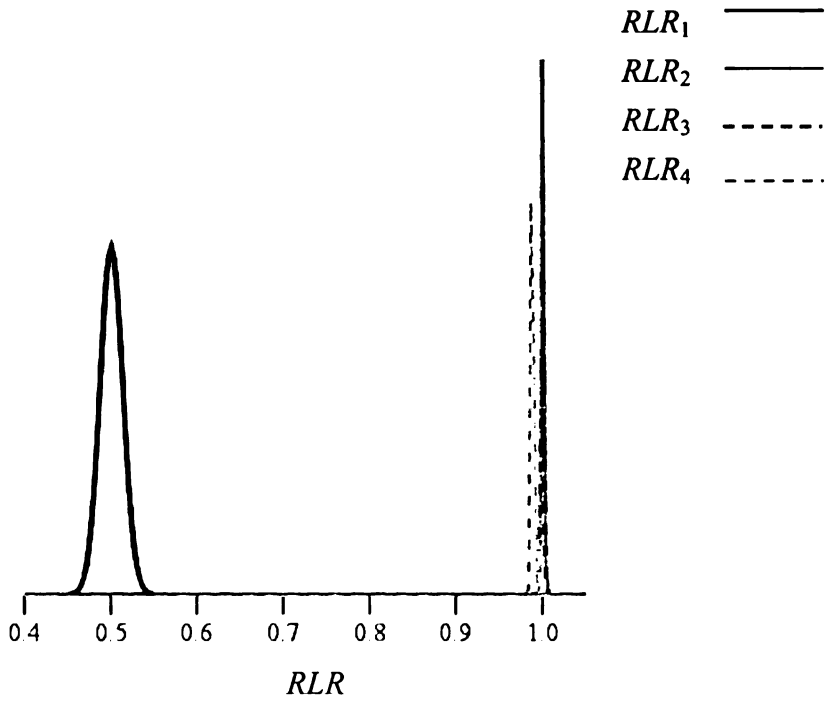


Figure C.14. The conditional distributions of the *RLR* values for Form 312

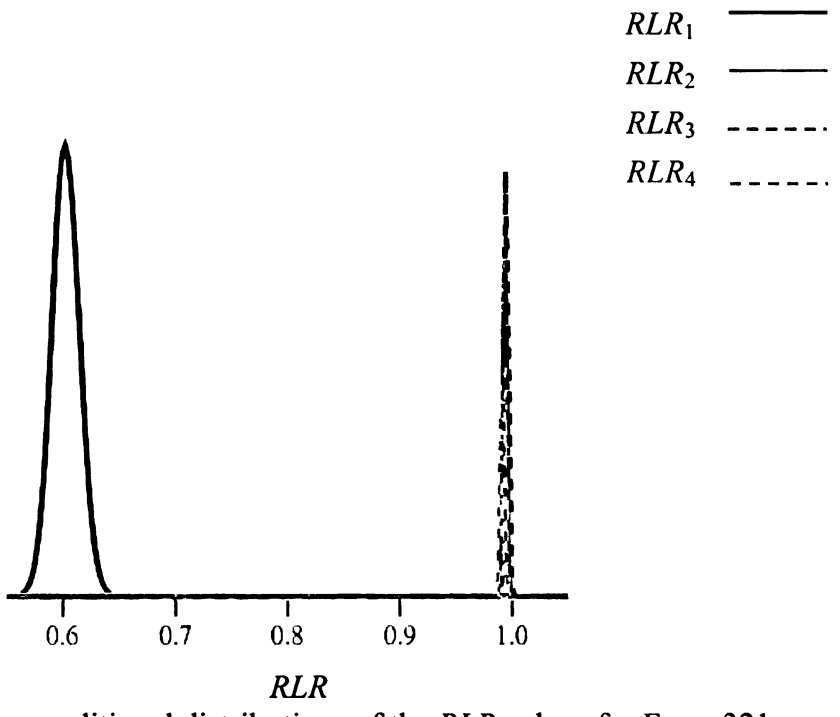


Figure C.15. The conditional distributions of the *RLR* values for Form 321

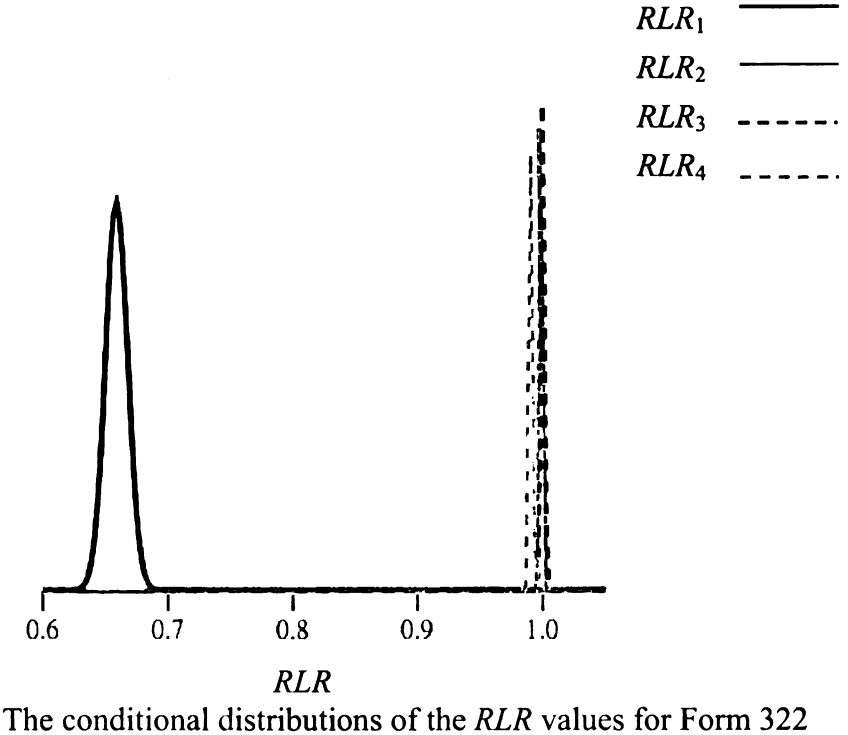


Figure C.16. The conditional distributions of the *RLR* values for Form 322

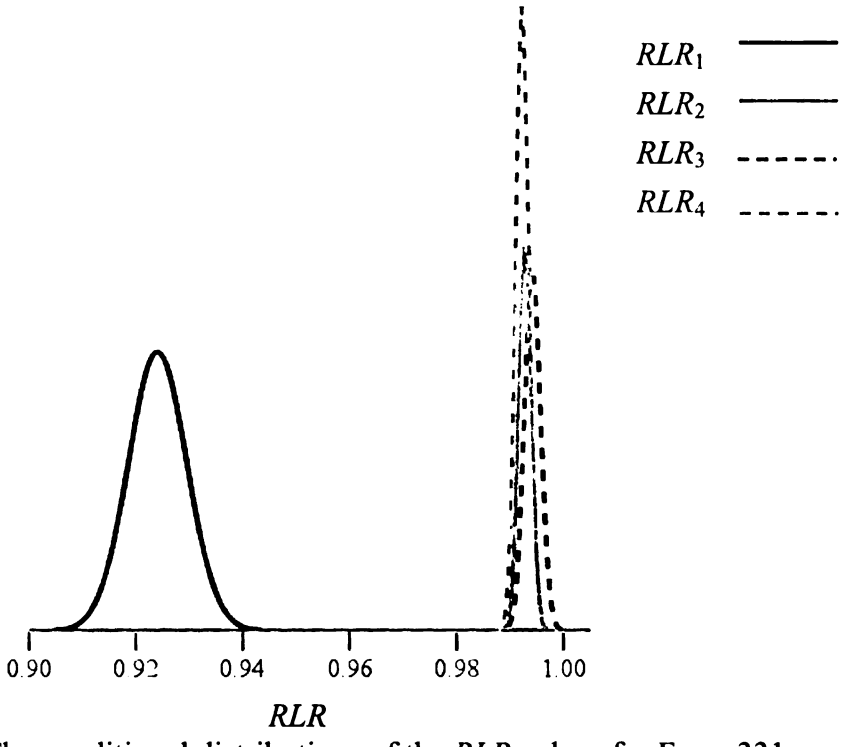


Figure C.17. The conditional distributions of the *RLR* values for Form 331

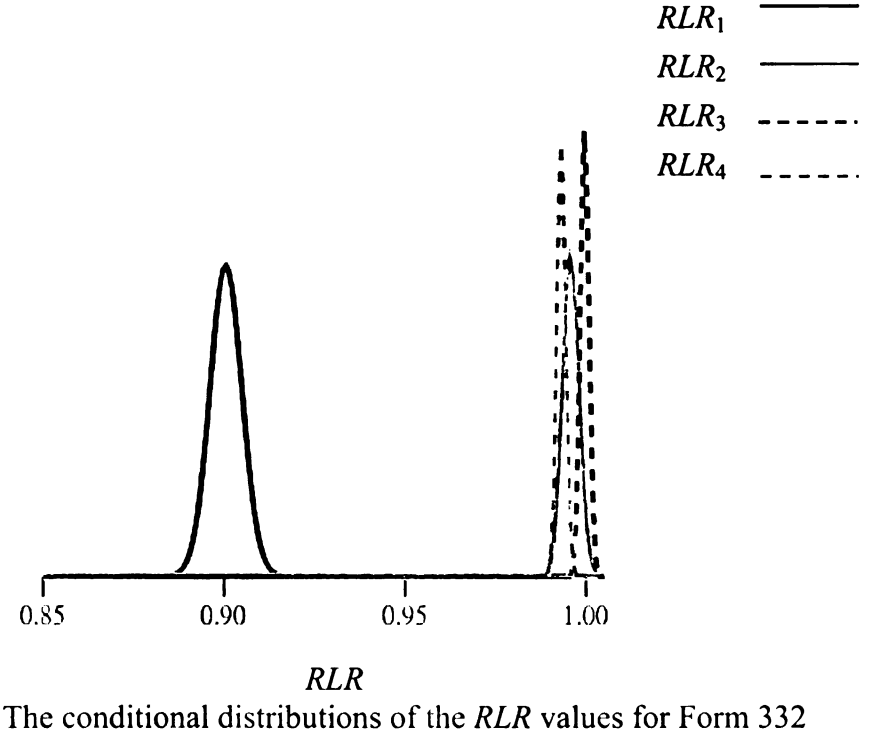


Figure C.18. The conditional distributions of the *RLR* values for Form 332

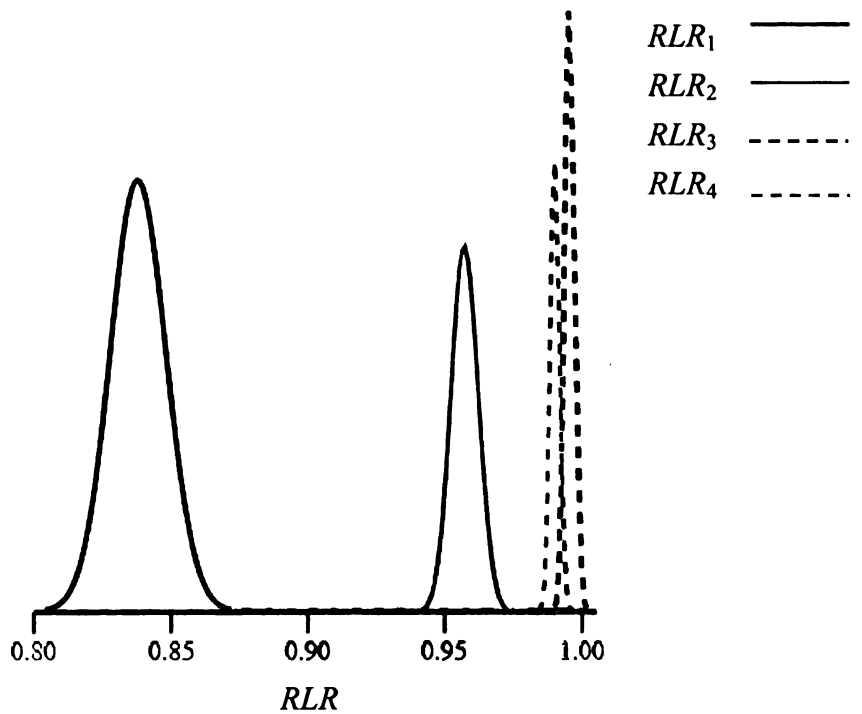


Figure C.19. The conditional distributions of the *RLR* values for Form 411

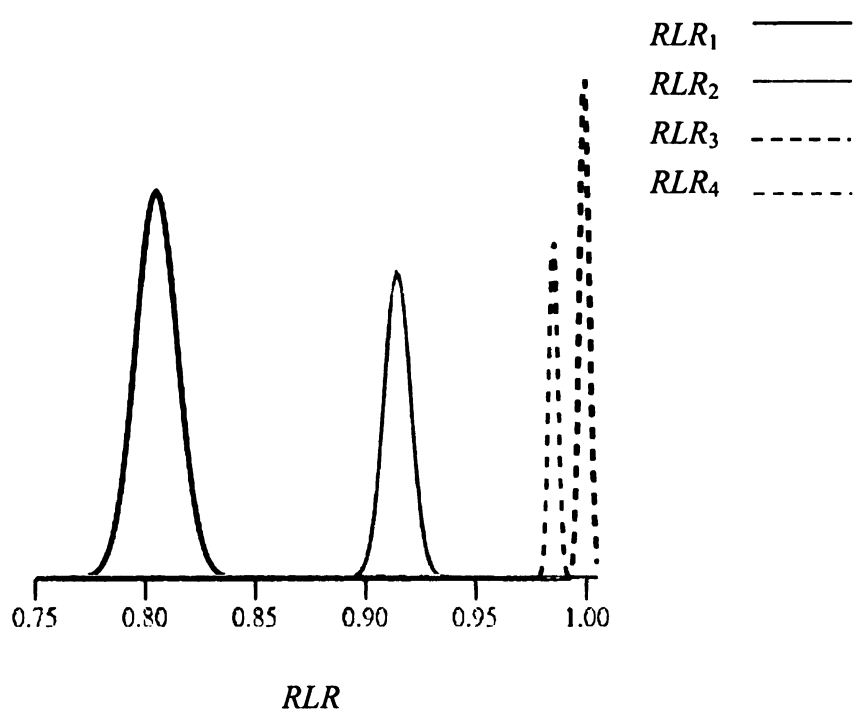


Figure C.20. The conditional distributions of the *RLR* values for Form 412

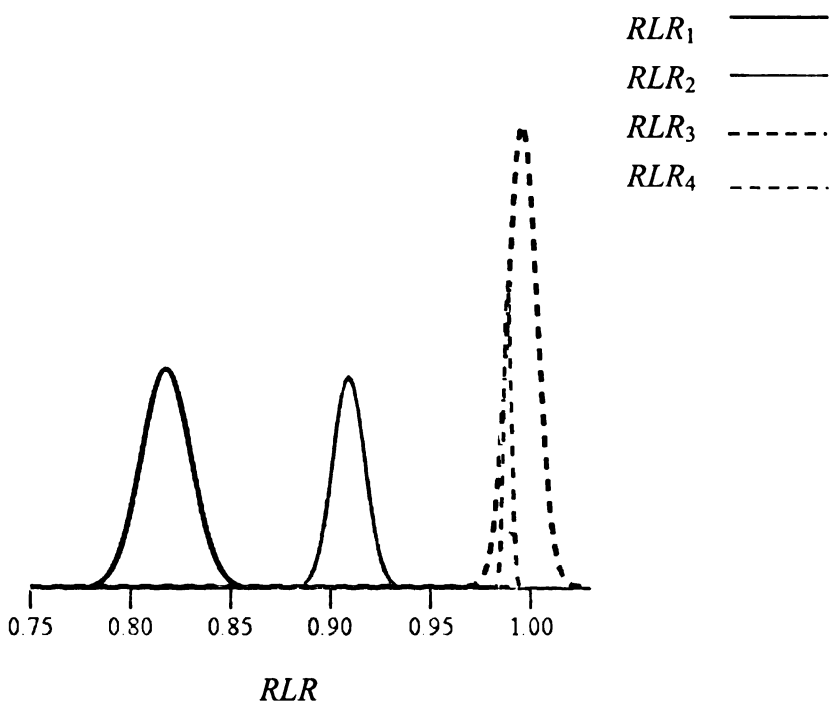


Figure C.21. The conditional distributions of the *RLR* values for Form 421

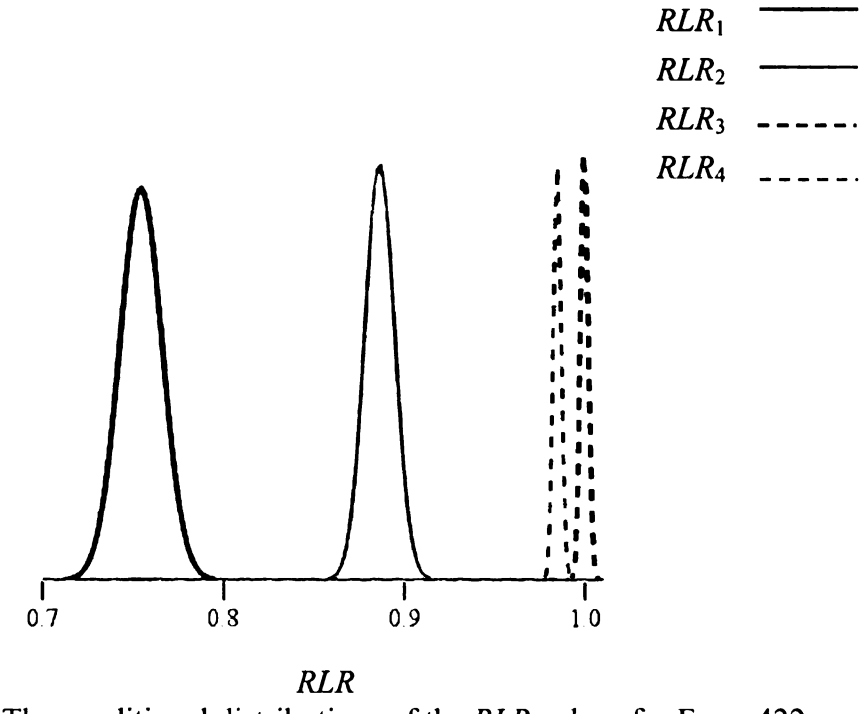


Figure C.22. The conditional distributions of the *RLR* values for Form 422

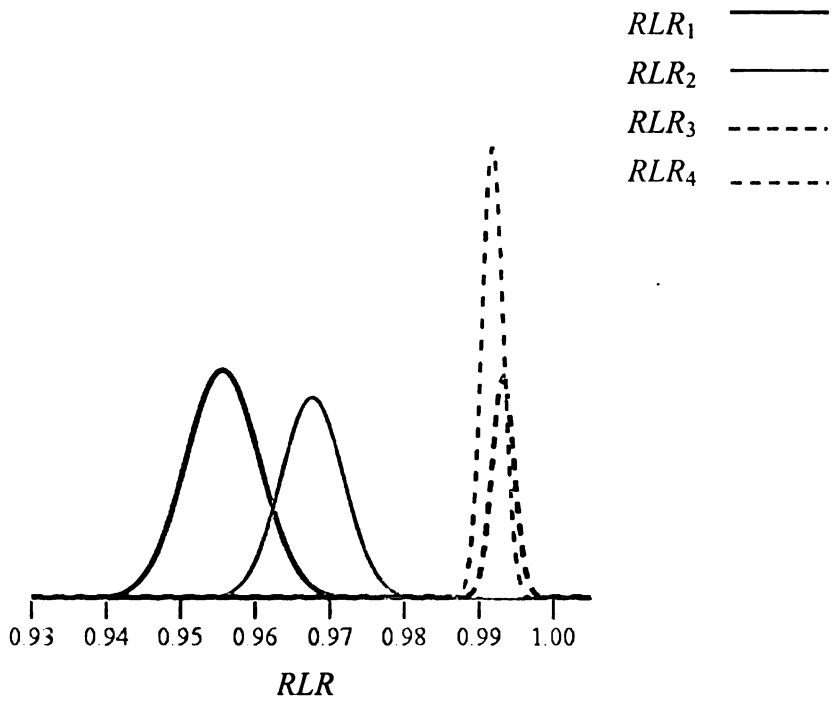


Figure C.23. The conditional distributions of the *RLR* values for Form 431

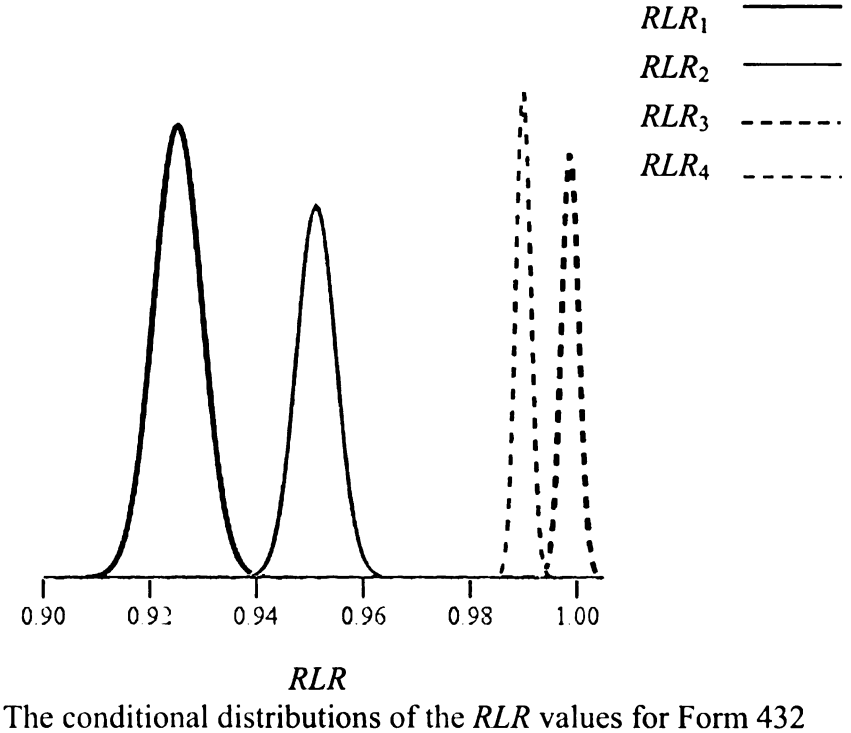


Figure C.24. The conditional distributions of the *RLR* values for Form 432

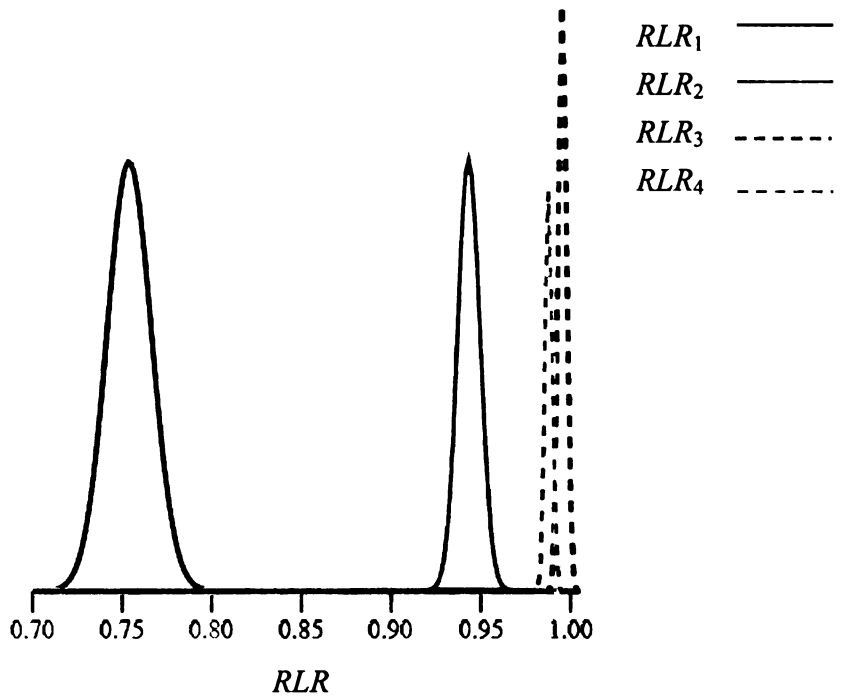


Figure C.25. The conditional distributions of the *RLR* values for Form 511

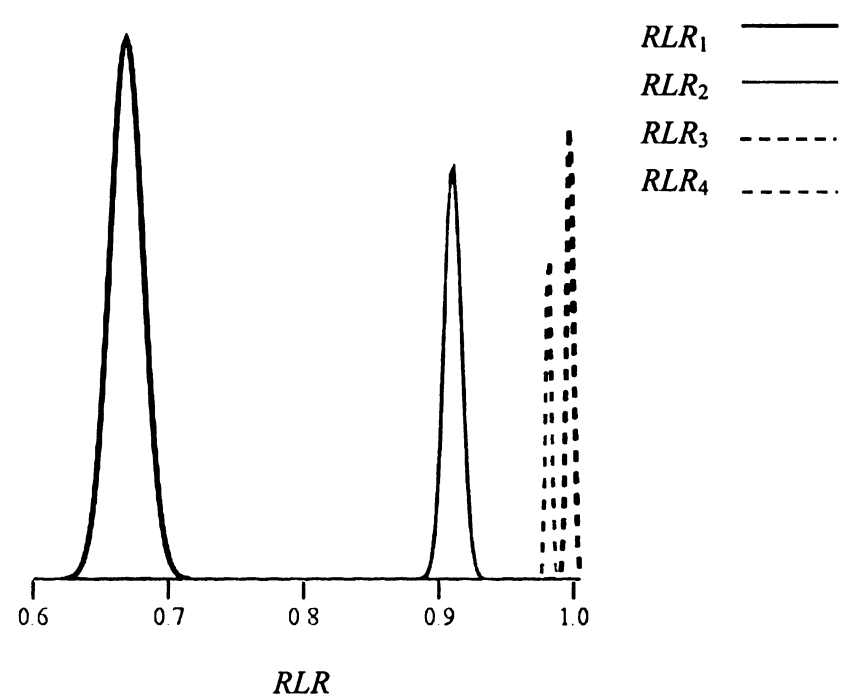


Figure C.26. The conditional distributions of the *RLR* values for Form 512

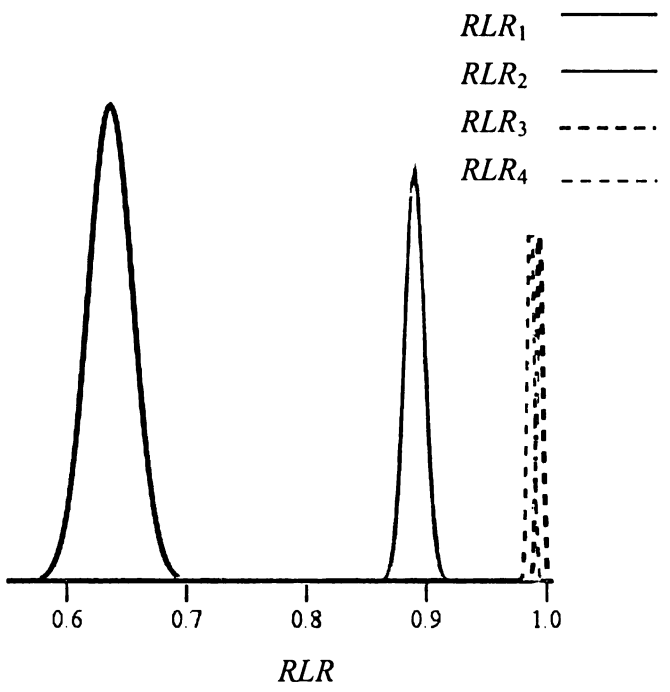


Figure C.27. The conditional distributions of the *RLR* values for Form 521

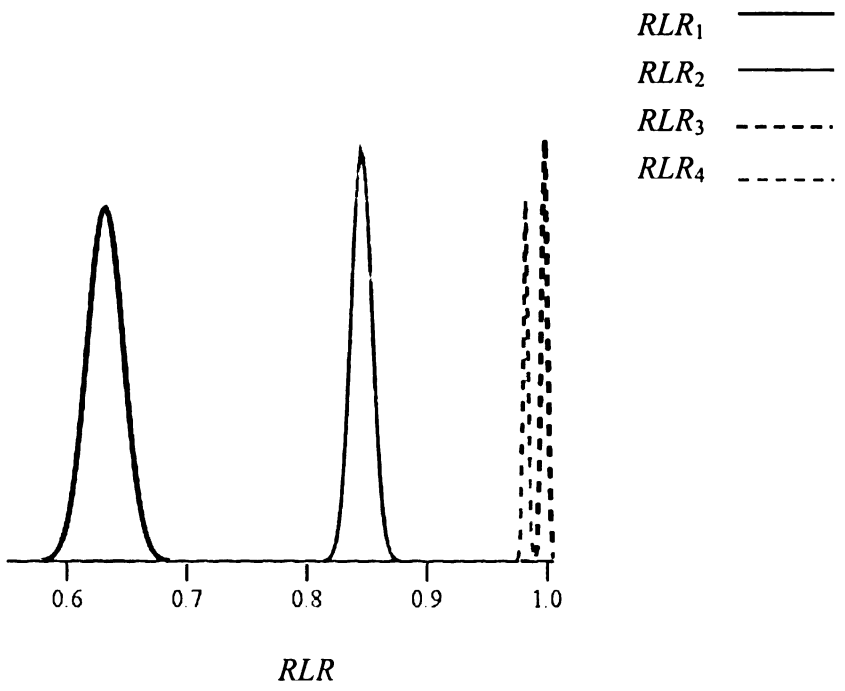


Figure C.28. The conditional distributions of the *RLR* values for Form 522

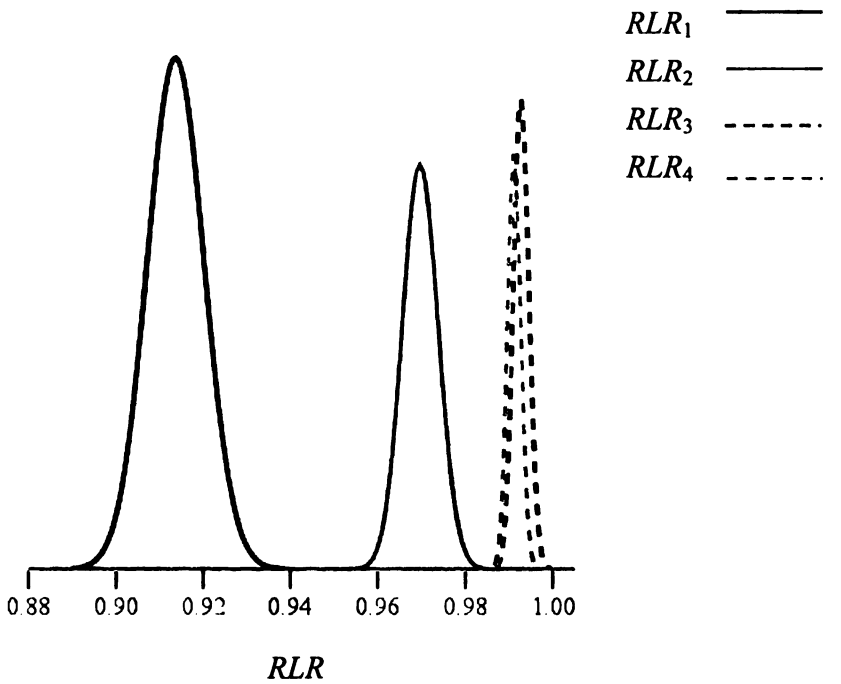


Figure C.29. The conditional distributions of the *RLR* values for Form 531

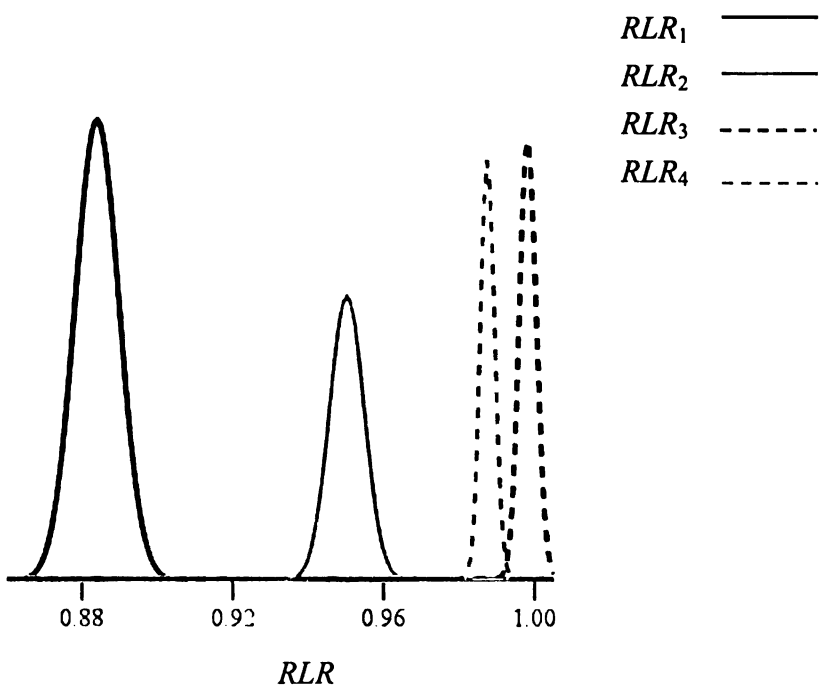


Figure C.30. The conditional distributions of the *RLR* values for Form 532

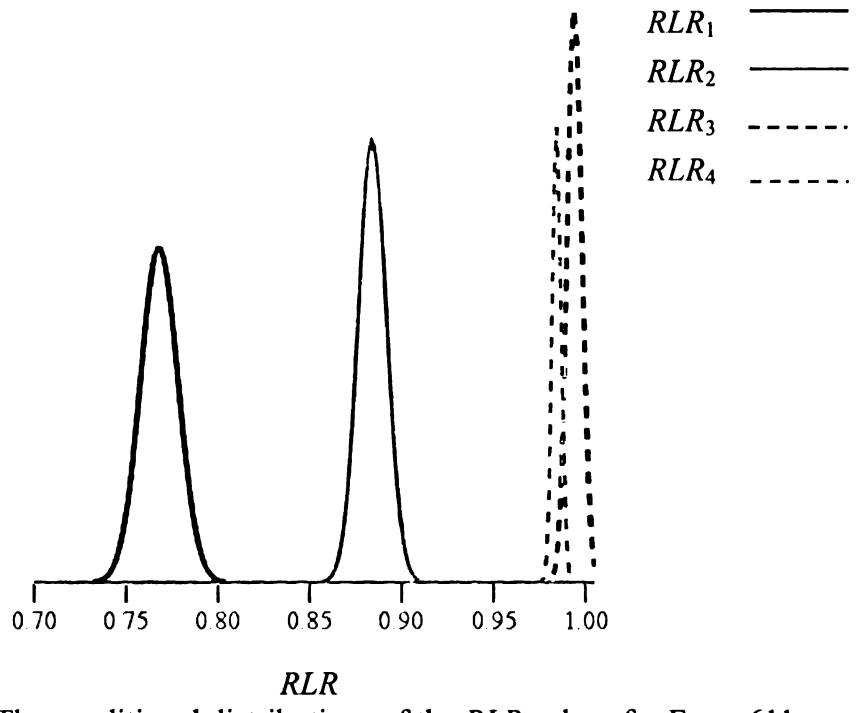


Figure C.31. The conditional distributions of the *RLR* values for Form 611

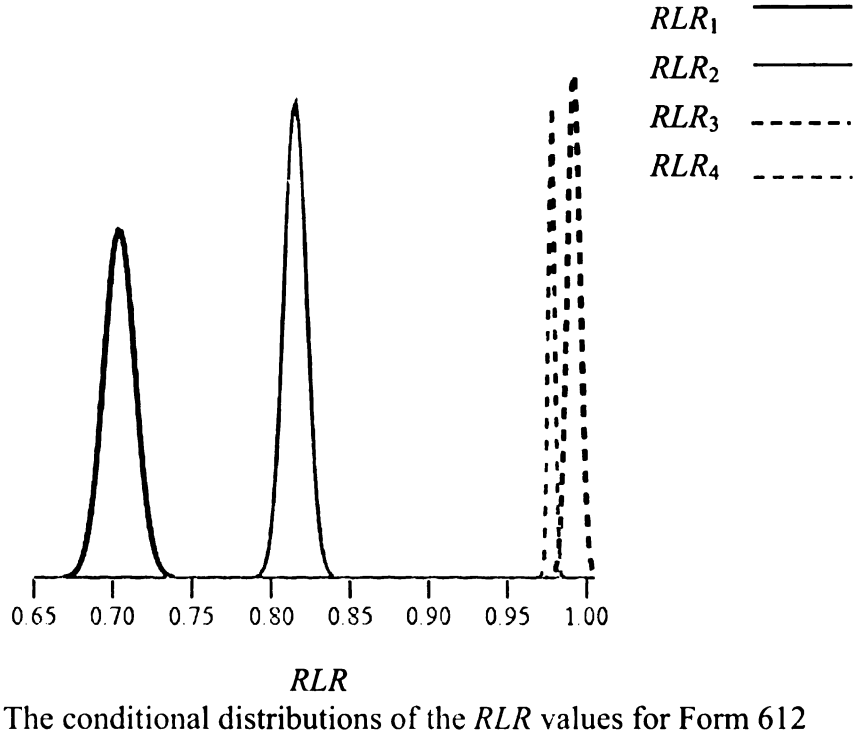


Figure C.32. The conditional distributions of the *RLR* values for Form 612

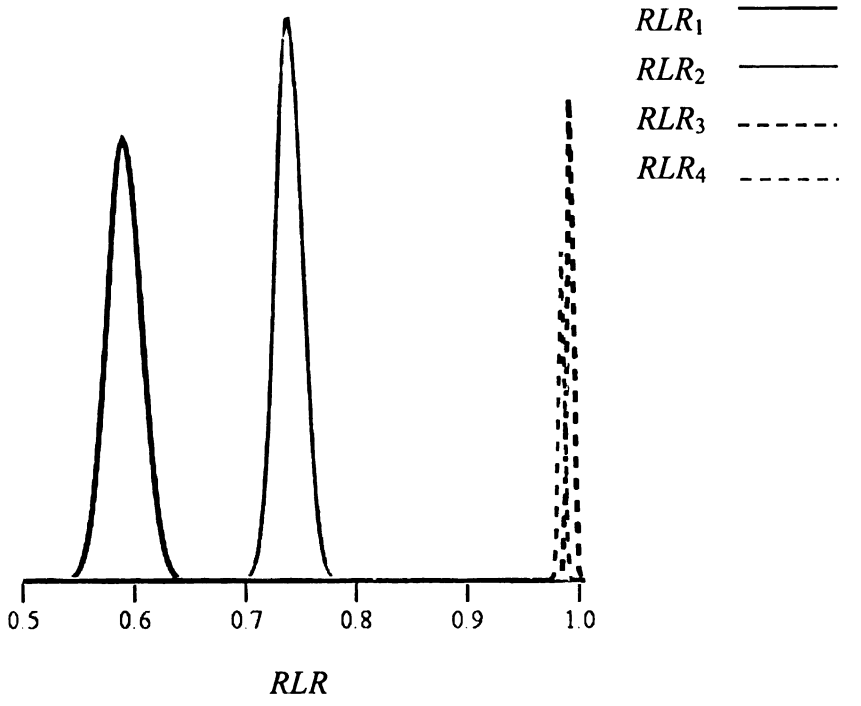


Figure C.33. The conditional distributions of the *RLR* values for Form 621

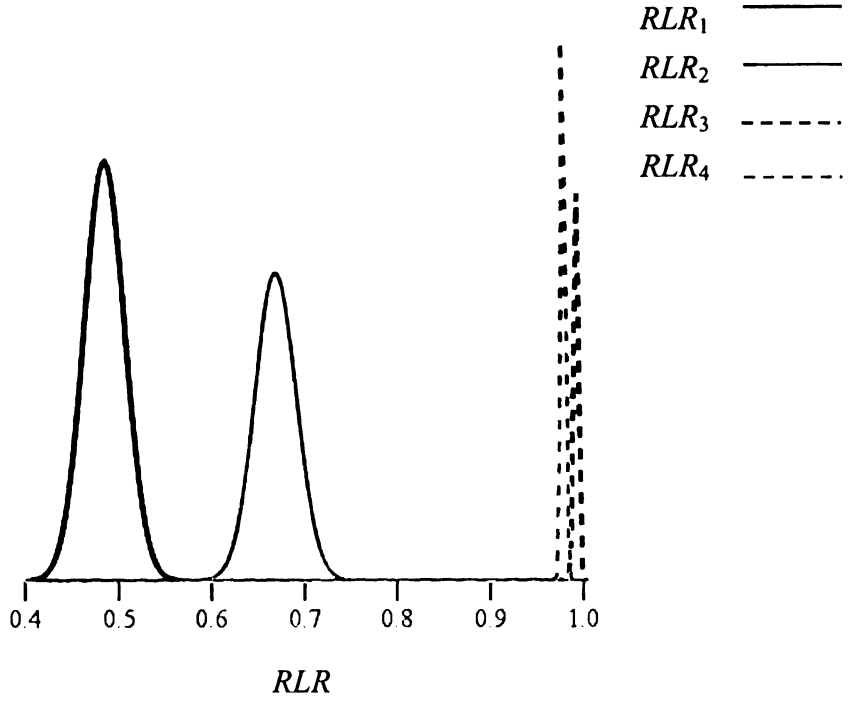


Figure C.34. The conditional distributions of the *RLR* values for Form 622

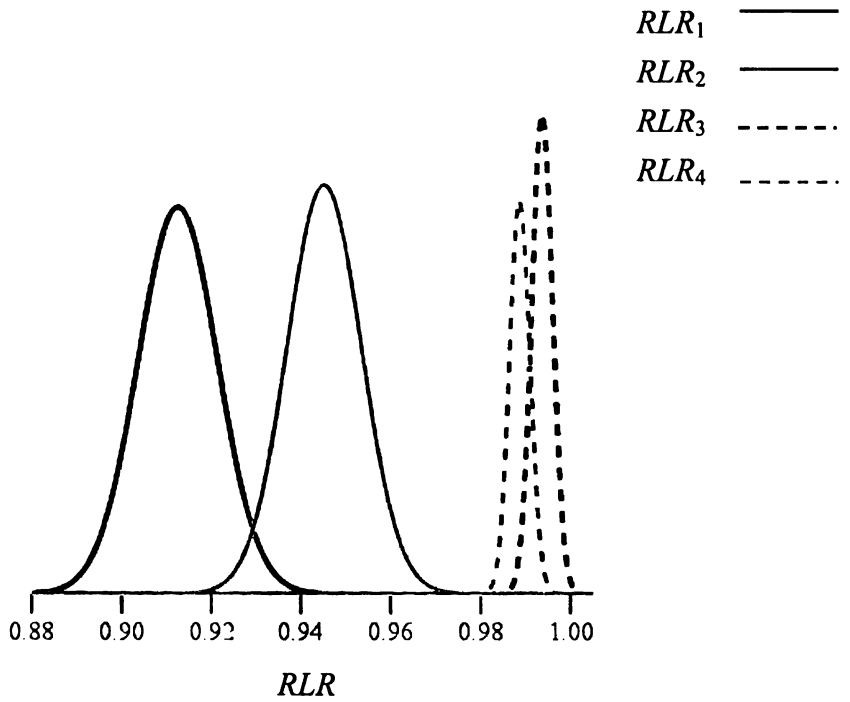


Figure C.35. The conditional distributions of the *RLR* values for Form 631

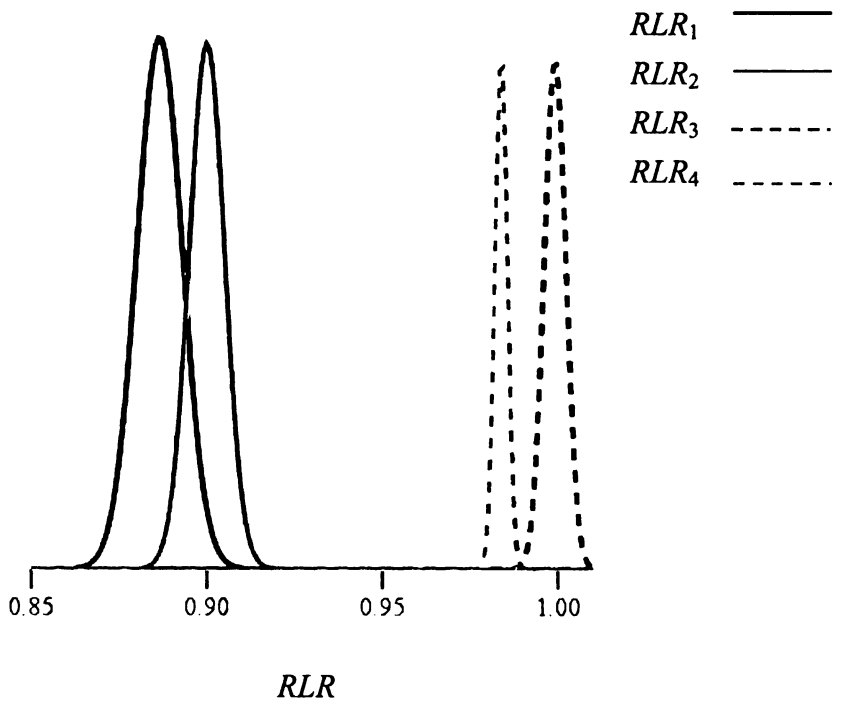


Figure C.36. The conditional distributions of the *RLR* values for Form 632

REFERENCES

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*(4), 255-278.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, 3*, 77-85.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17*, 283-296.

Bejar, I. I. (1988). An approach to assessing unidimensionality revisited. *Applied Psychological Measurement, 12*, 377-379.

Berger, M. P. F., & Knol, D. L. (1990). *On the assessment of dimensionality in multidimensional item response theory models. Research Report 90-8* (142 Reports--Evaluative). Netherlands: Twente University, Enschede (Netherlands). Department of Education.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information item factor analysis. *Applied Measurement in Education, 12*(3), 443-459.

Box, G. E. P. (1954). Some theorems on quadratic forms in the study of analysis of variance problem II. Effect of inequality of variance and correlation between errors in the two-way classification. *Annals of Mathematical Statistics, 25*, 484-498.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Choi. (1997). *A response dichotomization technique for item parameter estimation of the multidimensional graded response model.* Unpublished doctoral dissertation, University of Texas at Austin, Austin.

Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data* (No. 97-4). Iowa City, IA: College Admission Testing Program.

De Ayala, R. J., & Hertzog, M. A. (1991). The assessment of dimensionality for use in

item response theory. *Multivariate Behavioral Research, 26*(4), 765-792.

De Champlain, A., & Gessaroli, M. E. (1991). *Assessing test dimensionality using an index based on non-linear factor analysis.* Paper presented at the annual meeting of American Educational Research Association, Chicago, IL.

De Champlain, A., & Gessaroli, M. E. (1996). *Assessing the dimensionality of item response matrices with small sample sizes and short test lengths.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

De Champlain, A., & Gessaroli, M. E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. *Applied Measurement in Education, 11*(3), 231-253.

DeMars, C. E. (2003). Detecting multidimensionality due to curriculum differences. *Journal of Educational Measurement, 40*(1), 29-51.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics 23*(2), 129-151.

Douglas, J., Kim, H. R., Roussos, L., Stout, W., & Zhang, J. (1995). *LSAT dimensionality analysis for the December 1991, June 1992, and October 1992 administration*: LSAC research report series. LSAC-R-95-05.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*(3), 363-373.

Efron, B. (1978). Regression and ANOVA with zero-one data: measures of residual variation. *Journal of American Statistical Association, 73*, 113-121.

Embretson, S. E. (1985). Multicomponent latent trait models for test design. In S. E. Embretson (Ed.), *Test Design: Developments in Psychology and Psychometrics*. New York, NY: Academic Press.

Estrella, A. (1998). A new measure of fit for equation with dichotomous dependent variables. *Journal of Business & Economic Statistics, 16*(2), 198-205.

Estrella, A., Rodrigues, A. P., & Schich, S. (2003). How stable is the predictive power of the yield curve? Evidence from Germany and the United States. *The Review of*

Economics and Statistics, 85(3), 629-644.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large scale hands-on science performance assessment. *Journal of Educational Measurement 36*(2), 119-140.

Fraser, C. (1988). NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory. Center for Behavioral Studies. The University of New England. Armidale, New South Wales, Australia.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267-269.

Gessaroli, M. E., & De Champlain, A. (1996). Using an approximate x^2 statistics to test the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement, 33*, 157-179.

Guilford, J. P. (1941). The difficulty of a test and its factor composition. *Psychometrika, 6*, 67-77.

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected cells counts. *Annals of Statistics, 5*, 1148-1169.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*, 187-302.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11*(2), 91-115.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-125.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*(1), 49-78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Herath, P. H. M. U., & Takeya, H. (2003). Factors determining intercropping by rubber smallholders in Sri Lanka: a logit analysis. *Agricultural Economics, 29*(2), 159-168.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics, 14*, 1523-1543.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression.* New York, NY: John Wiely & Sons.

Howell, D. C. (2001). *Statistical Methods for Psychology* (fifth ed.). CA: Duxbury.

Humphreys, L. G. (1985). General intelligence: An integration of factor, test, and simplex theory. In B. B. Wolman (Ed.), *Handbook of intelligence* (pp. 201-224). New York, NY: Wiley.

Hutten, L. (1980). *Some empirical evidence for latent trait model selection.* Paper presented at the annual meeting of American Educational Research Association, Boston, MA.

Junker, B., & Stout, W. F. (1994). Robustness of ability estimation when multiple traits are presented with one trait dominant. In D. Laveault, B. D. Zumbo, M. E. Gessaroli & M. W. Boss (Eds.), *Modern Theories of Measurement: Problems and Issues* (pp. 31-61). Ottawa, Canada: Edumetrics Research Group, University of Ottawa.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20,* 141-151.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35,* 401-415.

Kendall, M. G. (1977). Multivariate contingency tables and for further problems in multivariate analysis. In P. R. Krishnaiah (Ed.), *Multivariate analysis IV.* Amsterdam: North Holland.

Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized test data* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Knol, D. L., & Berger, P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models *Multivariate Behavioral Research 26,* 457-477.

Kvalseth, T. O. (1985). Cautionary note about R^2. *The American Statistician, 39*(4), 279-285.

Lindman, H. R. (1974). *Analysis of variance in complex experimental designs.* New York, NY: W. H. Freeman.

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum.

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

Lumsden, J. (1957). A factorial approach to unidimensionality. *Australian Journal of Psychology, 9*, 105-111.

Magee, L. (1990). R^2 Measure based on Wald and likelihood ratio joint significance test. *The American Statistician, 44*(3), 250-253.

McDonald, R. P. (1967). Non-linear factor analysis. *Psychometric Monographs, 15*, (15, Pt. 12).

McDonald, R. P. (1981). The dimensionality of tests and items. *British journal of Mathematical and Statistical Psychology, 34*, 100-117.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*, 379-396.

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Lawrence Erlbaum.

McDonald, R. P. (1989a). Future directions of item response theory. *International Journal of Educational Research, 13*, 205-220.

McDonald, R. P. (1989b). A index of goodness-of-fit based on noncentrality. *Journal of classification, 6*, 97-103

McDonald, R. P., & Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research 30*(1), 23-40.

McKinley, R., L. (1989). *Confirmatory analysis of test structure using multidimensional item response theory* (No. ETS-RR-89-31). Princeton, NJ: Educational Testing Service.

McKinley, R., L., & Way, W. D. (1992). *The feasibility of modeling secondary TOEFL ability dimensions using multidimensional IRT models.* (No. ETS-RR-92-16). Princeton, NJ: Educational Testing Service.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54*(1), 17-24.

Michigan Department of Education. (2004, November 30, 2006). Mathematics grade level content expectation. Retrieved November 30, 2006, from http://www.michigan.gov/documents/ELA_K-8_87340_7.pdf

Michigan Department of Education. (2006). Mathematics Field Review. Retrieved

November 30, 2006, from http://www.michigan.gov/documents/mde/MATHEMATICS_Spreadsheet_092206_with_DUTCHER_BACKGROUND_AND_INSTRUCTIONS_174138_7.pdf

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*, 3-31.

Moneta, F. (2005). Does the yield spread predict recessions in the Euro Area? *International Finance, 8*(2), 263-301.

Muthen, B. (1987). LISCOMP: Analysis of linear structural equations using a comprehensive measurement model. Moorseville: Ind: Scientific Software.

Muthen, L. K., & Muthen, B. (1998). Mplus: The comprehensive modeling program for applied researchers. User's guide. Los Angeles: Muthen & Muthen.

Nandakumar, R. (1994). Assessing dimensionality of a set of responses-comparison of different approaches. *Journal of Educational Measurement, 31*(1), 17-35.

Nandakumar, R., & Stout, W. F. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18*(1), 41-68.

Olson, C. L. (1976). On choosing a test statistic in multivariate analyses of variance. *Psychological Bulletin, 83*, 579-586.

Reckase, M. D. (1985). The difficulty of test Items that measure more than one ability. *Applied Psychological Measurement, 9*(4), 401-412.

Reckase, M. D. (1990). *Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Reckase, M. D. (1997a). A linear logistic multidimensional model for dichotomous item response data. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York, NY: Springer.

Reckase, M. D. (1997b). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*(3), 193-203.

Reckase, M. D., & McKinley, R., L. (1982). *Some latent trait theory in a multidimensional latent space.* Paper presented at the Item Response Theory and Computerized Adaptive Testing Conference, Wayzata, MN.

Reckase, M. D., & McKinley, R., L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*(4), 361-373.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*(3), 425-435.

Roussos, L. A. (1992). Hierarchical agglomerative clustering computer program user's manual: University of Illinois at Urbana-Champaign.

Roussos, L. A. (1993). PROX help sheet. Urbana-Champaign: Statistical Laboratory for Educational and Psychological Measurement, Department of Statistics, University of Illinois.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35*(1), 1-30.

Roznowski, M., Tucker, L. R., & Humphreys, L. G. (1991). Three approaches to determining the dimensionality of binary items. *Applied Psychological Measurement, 15*(2), 109-127.

Shin, Y. S., & Moore, W. T. (2003). Explaining credit rating differences between Japanese and U.S. agencies *Review of Financial Economics, 12*(4), 327-344.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Steiger, J. H. (1980a). Testing pattern hypotheses on correlation matrices. *Multivariate Behavioral Research, 15*, 335-352.

Steiger, J. H. (1980b). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251.

Stone, C. A., & Yeh, C. C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the multistate bar examination. *Educational and Psychological measurement, 66*(2), 193-214.

Stout, W., Habing, B., Kim, J., Roussos, L., & Zhang, J. (1993). Conditional covariance based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*, 331-354.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617.

Stratmann, T. (2002). Can special interests buy congressional votes? Evidence from financial services legislation. *The Journal of Law and Economics, 45*, 345-373.

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? . *Synthese, 48*, 191-199.

Sympson, J. B. (1978). A model for testing with multidimensional items. In J. D. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference* (pp. 82-98). MN, Minneapolis: University of Minnesota.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & T. S. Long (Eds.), *Testing Structural Equation Models*. Newbury Park, CA: Sage.

Tate, R. (2002). Test dimensionality. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students*. Mahwah, NJ: Lawrence Erlbaum.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 23*(3), 159-203.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple category response models. *Journal of Educational and Behavioral Research, 26*, 247-260.

Thompson, T. (Undated). GENDAT5: A computer program for generating multidimensional item response data.

Turner, R. C. (2000). *Evaluating a procedure for investigating the multidimensional parallelism of standardized tests.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Turner, R. L., Miller, T., Reckase, M. D., Davey, T., & Ackerman, T. A. (1996). *Assessing the dimensionality of the interaction between items on a Mathematics test of the American College Testing (ACT) exam and subgroups of an ACT examinee population.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, New York, NY.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Vandaele, W. (1981). Wald, likelihood ratio, and Lagrange multiplier tests as an F test. *Economics Letters, 8*, 361-365.

Wainer, H., & Thissen, D. (1996). Howis reliability related to the quality of test scores? What is the effect of local item dependence on reliability? *Educational Measurement: Issues and Practice, 15*, 22-29.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479-494.

Wilson, D., Wood, R., Gibbons, R., Schilling, S., Muraki, E., & Bock, R. D. (2003). TESTFACT: Test scoring and full information item factor analysis (Version 4.0). Chicago, IL: Scientific Software International.

Zhang, J., & Stout, W. F. (1995). *Theoretical Results Concerning DETECT.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, San Francisco, CA.

Zhang, J., & Stout, W. F. (1996). *A new theoretical DETECT Index of dimensionality and its estimation.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, New York, NY.

Zheng, B., & Agresti, A. (2000). Summarizing the predictive power of a generalized linear model. *Statistics in Medicine, 19*(13), 1771-1781.