ED 422 394 TM 028 956

AUTHOR	Lee, Guemin; Kolen, Michael J.; Frisbie, David A.; Ankenmann, Robert D.
TITLE	Equating Test Forms Composed of Testlets Using Dichotomous and Polytomous IRT Models.
PUB DATE	1998-04-16
NOTE	42p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Equated Scores; *Item Response Theory; *Robustness (Statistics); Tables (Data); *Test Items
IDENTIFIERS	Dichotomous Variables; Polytomous Variables; *Testlets; Three Parameter Model

ABSTRACT

Item response models can be applied in many test equating situations by making strong statistical assumptions. Thus, studying the robustness of the models to violations of the assumptions and investigating model-data fit are essential in all item response theory (IRT) equating applications (M. Kolen and R. Brennan, 1995). Previous studies dealing with tests composed of testlets have indicated that the assumptions of the dichotomous IRT models are frequently violated; however, passage scores can be used instead of item scores to eliminate the effect of the dependence among within-passage items (G. Lee and D. Frisbie, 1997; H. Wainer and D. Thissen, 1996; S. Sireci, D. Thissen, and H. Wainer, 1991). The purpose of this study was to compare the performance of polytomous IRT models to that of the dichotomous three-parameter logistic model in the context of equating test forms composed of testlets, using traditional equating methods as criteria for both. For equating test forms composed of testlets, equating methods based on polytomous IRT models were found to produce results that more closely agreed with the results from traditional methods than did equating methods based on the dichotomous three-parameter logistic IRT model. (Contains 7 tables, 10 figures, and 32 references.) (Author/SLD)

Equating Test Forms Composed of Testlets Using Dichotomous and Polytomous IRT Models

Guemin Lee
Michael J. Kolen
David A. Frisbie
Robert D. Ankenmann
University of Iowa

Paper Presented at the 1998 Annual Meeting
of the National Council on Measurement in Education
San Diego, CA
April 16, 1998

Equating Test Forms Composed of Testlets Using Dichotomous and Polytomous IRT Models

Abstract

Item response models can be applied in many test equating situations by making strong statistical assumptions. Thus, studying the robustness of the models to violations of the assumptions and investigating model–data fit are essential in all IRT equating applications (Kolen & Brennan, 1995). Previous studies dealing with tests composed of testlets have indicated that the assumptions of the dichotomous IRT models are frequently violated; however, passage scores can be used instead of item scores to eliminate the effect of the dependence among within-passage items (Lee & Frisbie, 1997; Wainer & Thissen, 1996; Sireci, Thissen, & Wainer, 1991). The purpose of this study was to compare the performance of polytomous IRT models to that of the dichotomous three-parameter logistic model in the context of equating test forms composed of testlets, using traditional equating methods as criteria for both. For equating test forms composed of testlets, equating methods based on polytomous IRT models were found to produce results that more closely agreed with the results from traditional methods than did equating methods based on the dichotomous three-parameter logistic IRT model.

Equating Test Forms Composed of Testlets Using Dichotomous and Polytomous IRT Models

Introduction

When item response models are applied in test equating situations, strong statistical assumptions must be made – unidimensionality and local item independence. Because unidimensional dichotomous logistic item response models are frequently used for equating, it is important to study the robustness of these models to violations of the assumptions and to investigate model data fit (Kolen & Brennan, 1995).

This study deals with the application of item response theory (IRT) equating procedures to tests composed of testlets, small tests that are small enough to manipulate but large enough to carry their own context (Wainer & Kiely, 1987; Wainer & Lewis, 1990). Reading comprehension tests, containing sets of passages with collections of items, are examples of tests composed of testlets. Previous studies dealing with test scores obtained from tests composed of testlets have indicated that the local item independence assumption is likely to be violated, making it difficult to satisfy the unidimensionality assumption required by IRT modeling. That is, when several items in a test are related to a common passage or other common stimulus material, dependence is present among those items, meaning that conditional dependence exists (Wainer & Thissen, 1996; Lee & Frisbie, 1997; Yen, 1993). In this situation, the application of dichotomous IRT models to the equating of test forms composed of testlets might cause problems. Because there is little evidence in the literature about how the violation of IRT assumptions affects equating relationships involving testlets, it is not clear how serious the degree of distortion of equated scores might be.

When testlets are used, passage scores instead of item scores can be used to eliminate the influence of the dependence among within-passage items (Lee & Frisbie, 1997; Wainer & Thissen, 1996; Sireci, Thissen, & Wainer, 1991). Polytomous IRT models might be considered as alternatives to the dichotomous IRT models if this problem is serious. The purpose of this study is to investigate the feasibility of adopting various polytomous item response models in the context of equating test forms composed of testlets. The utility of these models is compared to that of the dichotomous IRT models by using traditional equating methods as criteria for both. The theoretical explanations of traditional and dichotomous IRT equating methods are presented in Kolen (1988), Cook & Eignor (1991), and Kolen & Brennan (1995). In this paper, background information is limited to the use of polytomous IRT models in equating test forms composed of testlets.

The multitude of polytomous item response models introduced during the last three decades includes Samejima's (1969) graded response model, Andrich's (1978) rating scale model, Master's (1982) partial credit model, and Bock's (1972) nominal model. With respect to testlet applications, Bock's nominal model has been used most often (Wainer & Thissen, 1996; Wainer, 1995; Sireci, Thissen, & Wainer, 1991; Wainer, Sireci, & Thissen, 1991) because "the testlet scores are nominal (or at most semi-ordered) responses...[because] a score of 1 may not always reflect higher proficiency than a score of 0, due to guessing" (Thissen, Steinberg, & Mooney, 1989, p.259). Although the graded response model is based on ordered response categories, its use in testlet-based equating applications may be appropriate. There would be an ordered quality to testlet-based scores if such scores corresponded to the extent of completeness of the examinee's reasoning process within a specific testlet. This *a priori* rationale seems to be reasonable with reading comprehension testlets, where several dichotomously-scored items relate to a single reading passage. The more of such items within a testlet that an examinee answers correctly, the more extensive his or her reasoning process. Therefore, in the present study, Samejima's (1969) graded response model is compared to Bock's (1972) nominal model with respect to performance in equating testlet-based test scores.

To apply polytomous item response models in this situation, testlet scores are obtained by summing the dichotomous item scores of the items that constitute the testlet. If testlet *j* consists of n_j items, the polytomous testlet score would be an integer between 0 and n_j, inclusive. In other words, a testlet consisting of n_j dichotomous items can be reconceptualized and treated as a single polytomous item having n_j+1 response categories. Each of the response categories (1, 2,

..., n_j+1) corresponds with one of the polytomous passage scores.

Under Bock's (1972) nominal model, the probability that an examinee with a given ability (θ) responds to category *k* in passage *j* is

$$P_{jk}(\theta) = \frac{\exp[a_{jk}\theta + c_{jk}]}{\sum_{k=1}^{K}\exp[a_{jk}\theta + c_{jk}]}, \tag{1}$$

where *j*=1,2,...,*J* (passages), *k*=1,2,...,*K* (categories). The constraints, $\sum_k a_{jk} = \sum_k c_{jk} = 0$, are imposed on this model. The parameters of this model are rescaled by using centered polynomials of the associated scores to represent the category-to-category changes in the a_k and b_k values: $a_{jk} = \sum_{p=1}^{P}\alpha_{jp}(k-\frac{K}{2})^p$ and $c_{jk} = \sum_{p=1}^{P}\gamma_{jp}(k-\frac{K}{2})^p$, where the parameters, $[\alpha_p, \gamma_p]_j$, p = 1,2, ..., P for $p \le K$, are the free parameters to be estimated from the data (Thissen, Steinberg, & Mooney, 1989). The a_{jk} and c_{jk} are parameters associated with the *k*th category of passage *j* that identify the shape of the testlet (or passage) category trace lines: a_{jk} is analogous to the discrimination parameter, and c_{jk} is analogous to the intercept parameter.

In this study, true score equating is used as one method. The true score is defined by

$$T(\theta) = \sum_{j=1}^{J}\sum_{k=1}^{K} u_{jk}P_{jk}(\theta), \tag{2}$$

where u_{jk} is a weight allocated to response category *k* of passage *j*, and all other symbols are as previously defined. The IRT true score equating method outlined by Cook & Eignor (1991) and Kolen & Brennan (1995) was applied in this situation.

To implement IRT observed score equating, the distribution of observed number-correct scores on each form must be obtained, and then the equipercentile method can be applied. In dichotomous IRT observed score equating, the compound binomial distribution can be used to generate the distribution of observed number-correct scores for examinees of a given ability (Lord & Wingersky, 1984). Hanson (1994) extended this algorithm, the so-called Lord &

Wingersky recursive formula, to polytomous items (Wang, Kolen & Harris, 1996):

For item 1, (3)

$P_1(X = x|\theta) = P(U_1 = x|\theta)$, x=0,1,2,...,n$_1$

For item k=2,3,4,...,K,

$$P_k(X = x|\theta) = \sum_{u=0}^{n_k} P_{k-1}(X = x - u)P(U_k = u|\theta),\ x=0,1,2,\ldots, \sum_{k=1}^{k} n_k,$$

where U_k represents a random variable for the score on item *k*, ranging from 0 to n_k.

After getting the observed number-correct distribution for examinees of a given ability, the observed score distribution of Form X (New Form) for examinees of various abilities can be found by accumulating the observed score distribution for examinees at each ability. If the distribution of ability is characterized by a discrete distribution on a finite number of equally spaced points, the observed score distribution for examinees of various abilities can be approximated by summing over abilities:

$$f(x) = \sum_{\theta} f(x|\theta)\psi(\theta), \qquad (4)$$

where $\psi(\theta)$ is the distribution of θ and $f(x|\theta)$ is the conditional number-correct score distribution given θ, which can be obtained by Equation 3. The observed score distribution of Form Y (Old Form), $g(y)$, can be found by using Equations 3 and 4 and replacing *x* with *y*. Then, the conventional equipercentile method can be used to find score equivalents (Kolen & Brennan, 1995; Zeng & Kolen, 1995).

Under Samejima's (1969) graded response model, consider passage *j* in which the number-correct score corresponding to the dichotomous items that constitute the passage can be classified into one of *K* categories, numbered 1 through *K* inclusive with consecutive integers, and "call such a response a 'graded response'..." (p.20). Then, the probability that a graded response to passage *j* is classified into category *k* or higher, given θ, is

$$P^*_{jk}(\theta) = \begin{cases} 1 & k = 1 \\ \dfrac{1}{1+\exp[-a_j(\theta - b_{j,k-1})]} & 2 \le k \le K \\ 0 & k > K \end{cases} \qquad (5)$$

The parameter a_j is the passage discrimination parameter, which is constant across the response categories of a particular passage (i.e., constant throughout the whole reasoning process). (This is another important way in which the graded response model differs from Bock's (1972) nominal model; in the nominal model the passage discrimination parameter is free to vary across the response categories of a particular passage.) The $b_{j,k-1}$ is the difficulty parameter of the category boundary *k*-1 ($2 \le k \le K$) for passage *j*, and it is free to vary among the category boundaries of a particular passage such that $b_{j,k-1} < b_{j,k}$. (Note that $b_{j,k-1}$ is the θ-value at which the probability of the response being classified into category *k* or higher is 0.5.) The probability that a graded response is classified in category *k*, given θ, is defined by $P_{jk}(\theta) = P^*_{jk}(\theta) - P^*_{j,k+1}(\theta)$, which is also written as

$$P_{jk}(\theta) = \begin{cases} 1 - \dfrac{1}{1+\exp[-a_j(\theta - b_{j1})]} & k = 1 \\ \dfrac{1}{1+\exp[-a_j(\theta - b_{j,k-1})]} - \dfrac{1}{1+\exp[-a_j(\theta - b_{jk})]} & 2 \le k \le K-1 \\ \dfrac{1}{1+\exp[-a_j(\theta - b_{j,k-1})]} & k = K \end{cases} \qquad (6)$$

An examinee's true score can be calculated by using Equations 2 and 6, and then the procedures for IRT true score equating can be applied. From Equations 3 and 6, the observed number-correct distribution for examinees of a given ability can be obtained, and the observed score distribution for examinees of various abilities can be found by Equation 4. Then, the procedures for IRT observed score equating can be applied to this situation.

Objectives

The objectives of this study were to:

1. Assess the local item dependence and dimensionality of tests composed of testlets to determine the appropriateness of the dichotomous IRT models and various polytomous IRT models in the context of equating these test forms.

2. Compare equating results from traditional equating methods (i.e., mean, linear, and equipercentile) with those from the dichotomous three-parameter logistic model and various polytomous item response models to investigate the implications of using these IRT models for equating test forms composed of testlets.

3. Investigate the generalizability of equating results for tests composed of various types of testlets, such as the Reading Comprehension, Maps and Diagrams, and Math Problem Solving and Data Interpretation tests of the Iowa Tests of Basic Skills (ITBS).

Method

Data Sources

The data for this study were taken from the 1995 ITBS Form M to Form K equating study. Data from the entire grade 8 sample of students for the Reading Comprehension (Reading), Maps and Diagrams (Maps), Math Problem Solving and Data Interpretation (Math), and Vocabulary tests were used (Hoover, Hieronymous, Frisbie & Dunbar, 1994). The sample size and the general characteristics of each test are presented in Table 1.

Insert Table 1 About Here

Though the Vocabulary tests do not have naturally-formed testlets, seven testlets were randomly formed for the purpose of comparison with tests composed of naturally-formed testlets. Two forms used in the equating, Form M (New Form) and Form K (Old Form), have exactly the same number of items per corresponding test.

Analyses

The local item independence assumption of the dichotomous IRT models was checked using Yen's (1984) Q_3 statistic, a correlation of the residuals of an item pair based on IRT models. The computer program IRT_LD (Chen, 1993; Chen & Thissen, 1997) was used to compute Yen's Q_3 statistic. The distributional characteristics of the pair of random variables (one for within-passage Q_3 and one for between-passage Q_3) for each form of each test were compared.

In order to check the unidimensionality assumption, principal component analyses and several exploratory factor analyses for each form of each test were completed. One analysis used a tetrachoric correlation matrix (obtained using PRELIS2) (Jöreskog & Sörbom, 1993) based on individual items, and the other used a product-moment correlation matrix based on testlet scores. Comparisons among eigenvalues from the principal component analyses were made and root mean squares were compared for each factor model.

The equating designs used in the 1995 ITBS Form M to Form K equating included a single group design and a random groups design. For tests used in this study, only a random groups design was used. Then, analyses were conducted with the RAGE (Zeng, Kolen & Hanson, 1995) computer program to find an equating function between both forms for each test using mean, linear, and equipercentile methods. For dichotomous IRT equating, item parameters were estimated by using the BILOG (Mislevy & Bock, 1990) program. (It was not necessary to place item parameters of the two forms on a common scale because a random groups design had been used for the equating.) True score and observed score equating relationships were found by using the PIE (Hanson & Zeng, 1995) computer program. The item parameters under Bock's nominal model and Samejima's graded response model were estimated with the MULTILOG (Thissen, 1991) program. The test characteristic curves for the tests in both forms and both true score and observed score equating functions were found by using a FORTRAN 90 program that was written for this purpose.

The moments of the distribution of equated scores from each equating method were calculated and compared. For comparing the overall level of discrepancy of each IRT equating method from the traditional equating methods, unweighted and weighted root mean squares (Harris & Crouse, 1993) were computed.

One restriction of the MULTILOG (Thissen, 1991) application program had to be addressed during the analysis: this program can accommodate at most 10 categories. However, both forms of the Reading test contained one passage that had more than 10 categories. In those cases, two categories were combined into a single category. Because the proportion of examinees in the combined categories was very small in both cases, the influence on the equating relationship of combining categories should not be significant in practical sense.

Results and Discussion

Local Independence

Yen's Q_3 statistic was used here as a measure of local item dependence. If there are n items in a test, $n(n-1)/2$ Q_3 statistics can be computed. In a similar way, for k_h items in the hth passage, there are $k_h(k_h-1)/2$ Q_3 statistics. Two types of Q_3 statistics were distinguished in this study for each form of each test: one is the within-passage Q_3 statistics (# of Q_3 = $\sum_{h=1}^{H} k_h(k_h-1)/2$), and the other is the between-passage Q_3 statistics (# of $Q_3 = n(n\text{-}1)/2 - \sum_{h=1}^{H} k_h(k_h-1)/2$). The distributional statistics for within-passage and between-passage Q_3 local item dependence measures are shown in Table 2.

--

Insert Table 2 About Here

--

Even though the Q_3 statistic is a correlation between residuals of an item pair based on IRT models (therefore, zero correlation might be expected for a locally independent item pair), Q_3 has a tendency to be slightly negative in the null case (Yen, 1984; Yen, 1993; Chen & Thissen, 1997). Yen (1993) demonstrated that the expected value of Q_3 statistics, when local independence is true, is approximately -1/(n-1), where n is the number of test items. These approximations of the expected values for the Q_3 statistics are also presented in Table 2. These values can be used as a criterion for comparing the overall level of local dependence of within- and between-passage item pairs.

The averages of the Q_3 statistics from within- and between-passage item pairs would be similar to the expected values of the Q_3 measures if the local item independence assumption holds. Table 2 shows that the averages of between-passage the Q_3 statistics for both forms of Reading, Maps, and Math tests have values similar to the expected values of Q_3 statistics, implying that item pairs between passages are locally independent. In contrast, the averages of within-passage Q_3 statistics for both forms of these tests have more positive values compared to the expected values of Q_3, even though the magnitudes of the differences in the Reading and Maps test forms are greater than in the Math test forms. This means that the local item independence assumption would be violated. For the Vocabulary test, because testlets were randomly constructed, averages of within- and between-passage Q_3 statistics are both similar to the expected value of Q_3, as would be anticipated. In comparing the difference between the observed mean and the expected mean of the Q_3 values with the standard deviation of the observed Q_3 statistics, in cases where local item dependence was identified, the magnitude of the difference seems to be about one standard deviation, except for the Math test forms. On the other hand, where local item dependence was not identified, the magnitude of the difference is much less than one standard deviation and close to zero.

Unidimensionality

Table 3 provides the first ten eigenvalues from tetrachoric correlation matrices based on individual items. These indicate that more than one factor would be required for explaining the data of both forms of the Reading and Maps tests and Form M of the Math test (at least the difference between the second eigenvalue and the third eigenvalue does not seem to be negligible compared to the difference between the third eigenvalue and the fourth eigenvalue). For both forms of the Vocabulary test and Form K of the Math test, one factor appears to be appropriate to explain the data. Scree plots for both forms of each test are presented in Figure 1.

Insert Table 3 About Here

Insert Figure 1 About Here

To get more information about the dimensionality of each form of each test, the root mean square (RMS) of the off-diagonal residuals under each specified number of factors was computed, as shown in Table 4. The difference between the RMSs of the one factor model and the two factor model from both forms of the Reading and Maps tests and Form M of the Math test are about two to six times greater than the difference between the RMSs of the two factor model and the three factor model. This means that one factor does not appear to be sufficient to describe the dimensionality of these test forms. For both forms of the Vocabulary test and Form K of the Math test, the difference between the RMSs of the one factor model and the two factor model is similar to the difference between the RMSs of the two factor model and the three factor model. Here, one factor seems sufficient to describe dimensionality. The results of several exploratory factor analyses, mainly comparing the RMSs, are consistent with the results from the principal component analyses.

Insert Table 4 About Here

On the basis of these results, it might be suspected that the unidimensional dichotomous IRT model might be problematic for equating test forms composed of testlets. That is, the common use of a unidimensional dichotomous IRT model in this equating situation could be suspect because of violations of assumptions. To check the possibility of adopting polytomous IRT models instead of using dichotomous IRT models, principal component analyses with product moment correlation matrices among testlet scores were conducted. Eigenvalues and scree plots are presented in Table 5 and Figure 2, respectively.

Insert Table 5 About Here

Insert Figure 2 About Here

One dominant factor is evident, and the other eigenvalues are considered negligible. The well-known Kaiser (1970) criterion, retaining eigenvalues greater than unity, has been criticized because of its susceptibility to the overidentification of dimensions (Cliff, 1988). Based on the Kaiser criterion, only one dimension is retained for all forms of all tests. In view of this susceptibility to overidentification, unidimensionality can be supported for the test forms used in this study, when testlet scores are used as the unit of analysis. Consequently, further analyses should not be needed in this situation, and the use of unidimensional polytomous item response models instead of dichotomous IRT models can be advocated on the basis of these results.

Comparisons with Traditional Equating Methods

Score conversions from mean, linear, equipercentile, dichotomous IRT true and observed score, Bock's nominal model true and observed score, and Samejima's graded response model true and observed score equatings for each test were tabulated and graphed. The results are complex and it is difficult to show the differences among equating functions because so many equating methods are displayed in one graph and table. Two general observations can be made, however:

1. All methods provide a similar equating relationship in the middle score range, but in the other score ranges it is reasonable to expect somewhat different equated scores from different equating methods.
2. All true and observed score equating methods of the IRT models (dichotomous IRT true and observed, Bock's nominal model true and observed, and Samejima's graded response model true and observed equating methods) produce similar equivalents to Form K from Form M along the score scale.

For a more convenient comparison among the various equating methods, difference scores can be used. The difference scores were calculated by subtracting the equated score of a baseline equating method (in this study, traditional equating methods such as the mean, linear, or equipercentile method) from the equated score of each equating method (in this study, dichotomous and polytomous IRT true or observed equating methods). To simplify the graph, the difference score plots of IRT true score and observed score equating methods are graphed separately. Even though it may be somewhat difficult to compare IRT true and observed score score equating methods by using separate graphs, there will be little loss of information because the IRT true and observed score equating methods provide similar equating relationships. On the other hand, because a main focus of this study is to compare the performance of dichotomous IRT models to that of polytomous item response models in equating test forms composed of testlets, the loss of information can be compensated for by providing the difference score plots of IRT true and observed equating methods separately. The plots for the Reading test are presented in Figures 3 and 4.

Insert Figure 3 About Here

Insert Figure 4 About Here

The vertical axis of the graphs in Figure 3 represents the difference score of each plotted equating method from the mean equating equivalents. One of the dotted lines, which is parallel to the horizontal axis and crosses the zero point on the vertical axis, represents a baseline equating method. The plotted line that was closest to this baseline represents the equating method that provides the equating function most similar to the referenced traditional equating method. The nominal and graded response model true score equating functions are much more similar to the mean and linear equating functions than the dichotomous IRT true score equating functions are. In the score range under 15, the nominal model true score equating method produces equivalents most similar to those of mean and linear equating. In the middle score range, from about 18 to 22, the equated scores of three methods are similar to those from both mean and linear equating methods. In the score range over 25, the graded response true score equating method provides equivalents most similar to those of mean and linear equating.

For the third graph in Figure 3, only for the scores of 30 and 31 does the dichotomous IRT true score equating method provide more similar equivalents to the baseline method. Otherwise, the nominal and graded response models provide more similar equivalents to those of the equipercentile equating method.

In Figure 4, which represents the difference scores between the IRT observed score equating methods and traditional equating methods, observed score equating methods based on polytomous IRT models provide more similar equating functions than the dichotomous IRT observed equating method does. For the first graph, in the score ranges under 13 and over 44, equated scores from the three methods are similar, but in the score range from 14 to 20, the dichotomous IRT observed score equating method is more similar, and on other score ranges (21-44), the graded response model is more similar. Similar trends can be found in the second graph using the linear equating method as a baseline. For the third graph, with the equipercentile equating method as a baseline, in the score range from 30 to 33, the dichotomous IRT observed score equating method provides more similar equivalents to those of the baseline method than do the other polytomous IRT observed equating methods. In the other score ranges, either the nominal or graded response model observed score equating method provides score equivalents more similar to those of the baseline equating method than does the dichotomous IRT observed score equating method.

As was found in the case of the Reading test, the polytomous IRT score equating methods, whether true score or observed score equating, give results that are much more similar to the referenced traditional equating methods than dichotomous IRT equating methods do for the Maps test. The difference score plots of both true and observed score equating methods for the Maps test are presented in Figures 5 and 6.

Insert Figure 5 About Here

Insert Figure 6 About Here

The main difference in the trends of the Maps test from those of the Reading test is that the graded response model equating methods provide similar equating functions in relation to the mean and linear equating methods. For the graphs using the equipercentile method as a baseline, the graded response model still provides the most similar equating function to that of the baseline method. Nominal model true or observed score equating results are more similar to the referenced traditional equating methods than are the dichotomous IRT equating true or observed score equating methods, except in the score ranges under 4 and over 26. Because the proportions of examinees in these score ranges (under 4 and over 26) are relatively small, the nominal model true and observed score equating methods would be expected to produce more similar equating functions to those of the traditional equating methods than do dichotomous IRT true or observed score equating methods.

The difference score plots of both true and observed score equating methods for the Math test are presented in Figures 7 and 8. Dichotomous IRT true score equating and polytomous IRT true score equating methods provide very similar equating functions, except in the score range under 11, where the polytomous IRT models are more similar to baseline methods. The similarity of the equating functions among these three methods is evident when comparing the observed score equating functions. The similarity among these three methods can be explained in terms of the violation of the assumptions for IRT modeling. That is, as previously indicated, the assumptions for dichotomous IRT modeling are less violated in the Math test compared to the Reading and Maps tests. So the similarity of the equating functions of the dichotomous and polytomous IRT models might not be surprising here. These results might be used as one piece of evidence to support the relationship between the degree of violation of IRT assumptions and its effect on equating relationships. Polytomous IRT true score equating methods provide slightly more similar equating relationships to the referenced traditional equating methods than does the dichotomous IRT true score equating method. This finding might be caused by

----

Insert Figure 7 About Here

Insert Figure 8 About Here

----

relatively large discrepancies within the lower portion of the score range. These relatively big discrepancies were not found in the observed score equating relationships. From these results, it might be hypothesized that IRT true score equating is more sensitive to the violation of the assumptions of IRT modeling than IRT observed score equating might be. This could be a topic for future research.

The Vocabulary test was included in this study for the purpose of comparison with tests composed of testlets because it can be thought of as the most unidimensional test in the ITBS test battery. Consequently, the unidimensional dichotomous IRT model could be expected to be the most appropriate model for equating test forms of this Vocabulary test. For comparing the results with other tests composed of testlets, 7 testlets were randomly constructed and the equating procedures of the polytomous IRT models were applied to find an equating function. Similar equating relationships would be expected to be found for both dichotomous and polytomous IRT equating methods. The difference score plots of both true and observed score equating methods for the Vocabulary test are presented in Figures 9 and 10.

Insert Figure 9 About Here

Insert Figure 10 About Here

At first, it might seem strange to find very different equating relationships among the equating methods. However, when the frequency distributions of the two forms of the Vocabulary test are examined, the unexpected result can be explained. That is, for Form M, there is no examinee with a score lower then 7, and only 5 percent of the examinees are under a score of 15. As a result, the item parameter estimation procedures are likely to be affected, especially in estimating the lower asymptote parameters. Compared to Form M, in Form K, 5 percent of the examinees are under a score of 10. The distributions of the two forms are very different. For this reason, these Vocabulary tests may not be good examples for use as a basis for comparison with tests composed of testlets. However, several important outcomes can be determined from the results of the equating of the Vocabulary test forms. First, the three true equating methods in Figure 9 provide similar equating functions, except in the score range under 17 (lower protion of the score scale). This finding makes sense if the frequency distributions of the two forms are considered. Second, ignoring the score range under 17, the dichotomous IRT true score equating method provides more similar equivalents on several score points, especially when using the equipercentile equating method as the baseline. Third, the similarity of the equating functions among the three methods are more evident in observed score equating than in true score equating. This finding is consistent with the conjecture made in analyzing the Math test.

Discussion so far has been based on the score conversions of dichotomous and polytomous IRT equating methods and their differences from baseline traditional methods. It would be also informative to summarize these differences using an overall index to represent the similarity and discrepancy of each equating method relative to the three baseline equating methods. For this purpose, Table 6 shows the moments for converted scores for each method and the absolute difference from the target, Form K moments in this case.

Insert Table 6 About Here

The mean of converted scores in Reading using the nominal model true and observed score equating methods are 24.82 and 24.78, respectively, and differences from the target are 0.03 and 0.07, respectively. These differences are much smaller than those of dichotomous IRT true and observed score equating methods (0.57 and 0.60, respectively). The nominal model also provides more similar standard deviation, skewness, and kurtosis values to the target than dichotomous IRT model true or observed score equating methods do. The graded response model true and observed score equating methods are also more similar than the dichotomous IRT model equating methods, in terms of their moments, even though they are less similar to the nominal model. In the Maps test, the graded response model provides much more similar moments to those of the target than do the other methods. The nominal model still provides more similar moments than the dichotomous IRT model does. For the Math test, the means of the nominal and graded response models are similar and are a little bit more similar to the mean of the target than that of dichotomous IRT model. However, in terms of standard deviation, skewness, and kurtosis, it is difficult to say which method provides more similar moments to those of the target. For the Vocabulary test, both true and observed score equating methods based on the graded response model have more similar means to those of the target, but have much different skewness and kurtosis values relative to those of the target than do dichotomous IRT true and observed score equating methods. As a result, it is hard to tell which method is more similar, in terms of moments, to those of the target. In short, for the Reading and Maps tests, the polytomous IRT model equating methods (either nominal or graded response model) produce more similar moments than the dichotomous IRT equating methods do, and for the Math test, they provide somewhat more similar moments. For the Vocabulary test, it is not possible to reach a general conclusion about the similarity of moments to the target among the various equating methods.

To provide a more direct overall index for comparing two equating methods, unweighted and weighted root mean squares were computed. The unweighted root mean square (URMS) is generally defined as $URMS = (\sum_i (A_i - B_i)^2 / k)^{1/2}$, where A_i is the equivalent of a raw score of i on the new test, B_i is another equivalent of a raw score of i on the new test, k represents the number of items, and i represents each raw score point. This URMS can be used to examine differences that occur throughout the score scale. However, this index does not take the score distribution of the new test (or distribution of equated scores) into account. The degree of distortion of the equated scores within a score range that includes a large proportion of examinees would be more important than the distortion of equated scores within a score range that includes a fairly small proportion of examinees. For this reason, the root mean square (RMS), which is an index weighted by the probability function of examinees at equated score points, is also computed by the formula of $RMS = (\sum_i f_i (A_i - B_i)^2 / \sum_i f_i)^{1/2}$, where f_i is the probability function of the raw score of the new form and the other notation is the same as for the

URMS. These two overall summary indices are presented in Table 7.

Insert Table 7 About Here

For the Reading test, using traditoinal equating methods as referents, polytomous IRT equating methods provide more similar equating relationships to the referenced methods than dichotomous IRT methods do. This finding is true when either the URMS or RMS is used, but the difference is more clear using the RMS than the URMS.

For the Maps test, the graded response model true or observed score equating methods are more similar to the three referenced methods than the methods based on dichotomous IRT or the nominal models, whether URMS or RMS is used. The differences in URMSs between the dichotomous IRT equating methods and the nominal model equating methods are not obvious, even though the differences in RMSs between the observed score equating methods based on the two models are more distinct (The RMS of the nominal model observed score equating method is smaller than the RMS of dichotomous IRT model observed score equating method is.).

For the Math test, according to the URMS, the differences among the three IRT models are not very great. However, the RMSs of the equating methods based on polytomous IRT models have somewhat smaller values than the RMSs of the equating methods based on the dichotomous IRT model. This result might be caused by the fact that the assumptions of dichotomous IRT modeling are violated less with the Math test than with the Reading or Maps tests. In other words, because the assumptions of dichotomous IRT modeling are violated less with the Math test, the degree of distortion of the equated scores is less severe than for the other tests composed of testlets.

For the Vocabulary test, the differences in the URMSs among equating methods based on each IRT model are not clearly distinct. Based on the RMSs, the graded response model provides more similar equating functions to the referenced equating methods, such as the mean and linear equating methods, than do other IRT models. One important observation is that the dichotomous IRT observed score equating method was found to be the best method when using the RMS as an index and the equipercentile method as a baseline. In sum, it is not possible to identify a single IRT model that consistently provides the most similar equating function to the referenced equating methods.

Conclusions

From the results presented, the polytomous IRT model true and observed score equating methods provide equating relationships that are more similar to traditional equating methods, such as mean, linear, and equipercentile methods, than do the dichotomous IRT true or observed score equating methods for the Reading, Maps, and Math tests. (This evidence is relatively less clear for the Math test than for the Reading and Maps tests.) The reason might be explained by the violation of the assumptions of dichotomous IRT modeling – the unidimensionality and local item independence assumptions – when testlets are used. A comparison of the equating functions derived from traditional methods with those obtained with IRT methods is one possible way to check the validity of IRT–model equating methods. Because polytomous IRT models satisfy the assumptions of IRT modeling more closely than do dichotomous IRT models for the case of tests composed of testlets, it is reasonable to expect better equating relationships using polytomous IRT equating methods than using dichotomous IRT equating methods. The nominal model and graded response model seem to offer the best alternatives for equating test forms composed of testlets.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Chen, W. (1993). *IRT_LD: A computer program for the detection of pairwise local dependence between test items* (Research Memorandum 93-2). Chapel Hill: L.L. Thurstone Laboratory, University of North Carolina at Chapel Hill.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289.

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin, 103*, 276-279.

Cook, L.L., & Eignor, D.R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice, 10,* 37-45.

Hanson, B.A. (1994). *An extension of the Lord-Wingersky algorithm to the polytomous items.* Unpublished research note.

Hanson, B.A., & Zeng, L. (1995). *A computer program for IRT equating (PIE) (version 1.0).* Iowa City, IA: ACT.

Harris, D.J., & Crouse, J.D. (1993). A study of criterion used in equating. *Applied Measurement in Education, 6*(3), 195-240.

Hoover, H.D., Hieronymous, A.N., Frisbie, D.A., & Dunbar, S.B. (1994) *Iowa tests of basic skills : Interpretive guide for school administrators.* Chicago, IL: The Riverside Publishing Company.

Jöreskog, K.G., & Sörbom, D. (1993). *PRELIS2 User's reference guide.* Chicago, IL: Scientific Software International, Inc.

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

Kolen, M.J. (1988). An NCME instructional module on traditional equating methodology. *Educational Measurement: Issues and Practice, 7,* 29-36.

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

Lee, G., & Frisbie, D.A. (1997, March). *A generalizability approach to evaluating the reliability of testlet-based scores.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Mislevy, R.J., & Bock, R.D. (1990). *BILOG 3. Item analysis and test scoring with binary logistic models (2nd ed.).* Mooresville, IN: Scientific Software.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement, No. 17.*

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

Thissen, D. (1991). *MULTILOG Multiple categorical item analysis and test scoring using item response theory (version 6.0).* Chicago, IL: Scientific Software.

Thissen, D., Steinberg, L., & Mooney, J.A. (1989). Trace lines for testlets: A use of multiple-categorical models. *Journal of Educational Measurement, 26*(3), 247-260.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. *Applied Measurement in Education, 8,* 157-186.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing : A case for testlets. *Journal of Educational Measurement, 24*(3), 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*(1), 1-14.

Wainer, H., Sireci, S.G., & Thissen, D. (1991). Differential testlet functioning: Definition and detecting. *Journal of Educational Measurement, 28*, 197-219.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22-29.

Wang, T., Kolen, M.J., & Harris, D.J. (1996). *Conditional standard errors, reliability and decision consistency of performance levels using polytomous IRT*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125-145.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213.

Zeng, L., & Kolen, M.J. (1995). An alternative approach for IRT observed-score equating of number-correct scores. *Applied Psychological Measurement, 19*(3), 231-240.

Zeng, L., Kolen, M.J., & Hanson, B.A. (1995). *Random groups equating program (RAGE) (version 2.0)*, Iowa City, IA: ACT.

TABLE 1
Descriptive Statistics for Data Sources Used in This Study

Characteristic	Reading	Maps	Math	Vocabulary
ITBS Form K (Old Form)				
Sample Size	663	632	537	666
No. of Items	49	33	36	43
No. of Passages	8	5	8	7
No. of Items per Passage	9,4,7,5,5,6,3,10	7,7,6,6,7	8,4,4,4,4,4,4,4	7,6,6,6,6,6,6
$\overline{X}$	24.9	16.3	16.5	24.4
S_X	9.99	6.34	6.38	8.87
Skewness	0.375	0.383	0.363	0.018
Kurtosis	2.216	2.306	2.413	2.170
ITBS Form M (New Form)				
Sample Size	680	653	561	680
No. of Items	49	33	36	43
No. of Passages	7	5	7	7
No. of Items per Passage	8,4,9,4,5,8,11	7,7,6,6,7	8,6,6,4,4,4,4	7,6,6,6,6,6,6
$\overline{X}$	25.9	15.3	19.2	27.1
S_X	10.53	6.31	6.67	6.88
Skewness	0.235	0.408	0.191	-0.244
Kurtosis	2.065	2.444	2.257	2.623

Note : Reading = Reading Comprehension, Maps = Maps and Diagrams, Math = Math Problem Solving and Data Interpretation

TABLE 2

Distribution Yen's Q_3 Statistics for Within-Passage and Between-Passage Item Pairs

Test	No. of Q_3	E (Q_3)	Mean	Diff.	S.D.	Skewness	Kurtosis	Range
Reading (K)	1176	-.021						
Between	1030		-.021	.000	.043	.006	3.045	-.180–.121
Within	146		.038	.059	.058	.044	3.328	-.131–.196
Reading (M)	1176	-.021						
Between	1007		-.027	.006	.045	-.103	3.016	-.181–.111
Within	169		.058	.079	.069	1.014	6.097	-.080–.400
Maps (K)	528	-.031						
Between	435		-.029	.002	.045	-.056	2.838	-.177–.098
Within	93		.031	.062	.049	.375	2.886	-.064–.164
Maps (M)	528	-.031						
Between	435		-.033	.002	.046	.166	3.002	-.167–.177
Within	93		.030	.061	.070	1.018	4.905	-.118–.281
Math (K)	630	-.029						
Between	560		-.022	.007	.049	-.013	3.084	-.183–.137
Within	70		.008	.037	.057	.525	2.858	-.087–.166
Math (M)	630	-.029						
Between	548		-.024	.005	.045	.028	3.581	-.159–.186
Within	82		.002	.031	.056	.008	4.440	-.191–.172
Vocabulary (K)	903	-.024						
Between	792		-.018	.006	.044	-.098	2.833	-.147–.106
Within	111		-.012	.012	.038	-.308	2.431	-.103–.059
Vocabulary (M)	903	-.024						
Between	792		-.016	.008	.044	-.176	3.078	-.174–.135
Within	111		-.018	.006	.044	-.057	3.010	-.120–.116

Notes : Reading = Reading Comprehension, Maps = Maps and Diagrams, Math = Math Problem Solving and Data Interpretation; E (Q_3) = Expected value of Q_3, Diff. = Absolute value of the difference between E (Q_3) and the sample mean.

TABLE 3
First Ten Eigenvalues of Tetrachoric Correlation Matrices Based on Individual Item Scores

Rank of	Reading (49 Items)		Maps (33 Items)		Math (36 Items)		Vocabulary (43 Items)	
Eigenvalue	Eigenvalue	Difference	Eigenvalue	Difference	Eigenvalue	Difference	Eigenvalue	Difference
ITBS Form K (Old Form)								
1	14.07	11.15	7.84	6.00	8.41	6.42	13.59	11.92
2	2.92	1.23	1.84	0.34	1.99	0.19	1.67	0.19
3	1.68	0.17	1.50	0.07	1.80	0.34	1.48	0.07
4	1.51	0.19	1.43	0.11	1.46	0.11	1.41	0.15
5	1.32	0.01	1.32	0.08	1.35	0.07	1.26	0.06
6	1.31	0.06	1.24	0.05	1.28	0.05	1.20	0.01
7	1.25	0.05	1.19	0.04	1.23	0.04	1.19	0.03
8	1.20	0.02	1.15	0.09	1.19	0.05	1.16	0.06
9	1.18	0.03	1.06	0.01	1.14	0.07	1.10	0.03
10	1.15	0.03	1.05	0.05	1.07	0.02	1.07	0.02
ITBS Form M (New Form)								
1	15.83	12.60	7.65	5.59	9.13	7.08	10.77	8.73
2	3.23	1.56	2.06	0.57	2.05	0.49	2.04	0.31
3	1.67	0.25	1.49	0.04	1.56	0.08	1.73	0.15
4	1.42	0.09	1.45	0.04	1.48	0.18	1.58	0.11
5	1.33	0.02	1.41	0.17	1.30	0.04	1.47	0.10
6	1.31	0.06	1.24	0.08	1.26	0.01	1.37	0.07
7	1.25	0.06	1.16	0.05	1.25	0.11	1.30	0.01
8	1.19	0.05	1.11	0.02	1.14	0.01	1.29	0.10
9	1.14	0.05	1.08	0.04	1.13	0.07	1.19	0.07
10	1.09	0.04	1.04	0.05	1.06	0.02	1.12	0.01

Note : Reading = Reading Comprehension, Maps = Maps and Diagrams, Math = Math Problem Solving and Data Interpretation

TABLE 4
Root Mean Square of Off-diagonal Residuals for Specified Numbers of Factors

Number of	Reading (49 Items)		Maps (33 Items)		Math (36 Items)		Vocabulary (43 Items)	
Factors	RMS	Difference	RMS	Difference	RMS	Difference	RMS	Difference
				ITBS Form K (Old Form)				
1	7.2	1.3	6.8	1.0	7.1	0.8	5.7	0.5
2	5.5	0.4	5.8	0.5	6.3	0.7	5.2	0.3
3	5.1	0.3	5.3	0.4	5.6	0.5	4.9	0.3
4	4.8	0.2	4.9	0.4	5.1	0.3	4.6	0.2
5	4.6	0.3	4.5	0.3	4.8	0.3	4.4	0.3
6	4.3	0.3	4.2	0.3	4.5	0.3	4.1	0.2
7	4.1	0.2	3.9	0.4	4.2	0.3	3.9	0.3
8	3.9	0.2	3.5	0.2	3.9	0.3	3.6	0.2
9	3.7	0.2	3.3	0.3	3.6	0.2	3.4	0.2
10	3.5		3.0		3.4		3.2	
				ITBS Form M (New Form)				
1	7.7	2.3	7.3	1.3	7.1	1.1	6.9	0.6
2	5.4	0.4	6.0	0.5	6.0	0.5	6.3	0.4
3	5.0	0.3	5.5	0.5	5.5	0.4	5.9	0.4
4	4.7	0.3	5.0	0.5	5.1	0.4	5.5	0.4
5	4.4	0.2	4.5	0.3	4.7	0.3	5.1	0.3
6	4.2	0.3	4.2	0.3	4.4	0.3	4.8	0.2
7	3.9	0.2	3.9	0.3	4.1	0.3	4.6	0.3
8	3.7	0.2	3.6	0.3	3.8	0.2	4.3	0.3
9	3.5	0.1	3.3	0.3	3.6	0.3	4.0	0.2
10	3.4		3.0		3.3		3.8	

Note : Reading = Reading Comprehension, Maps = Maps and Diagrams, Math = Math Problem Solving and Data Interpretation, RMS = root mean square of off-diagonal residuals. The scales of the RMS and difference have been changed by multiplying all entries by 100 and then rounding to one decimal place.

TABLE 5
Eigenvalues of Product-Moment Correlation Matrices Based on Passage Scores

Rank of	Reading		Maps		Math		Vocabulary	
Eigenvalue	Eigenvalue	Difference	Eigenvalue	Difference	Eigenvalue	Difference	Eigenvalue	Difference
				ITBS Form K (Old Form)				
1	4.10	3.30	2.59	1.91	3.34	2.51	4.44	3.95
2	0.80	0.17	0.68	0.06	0.83	0.01	0.49	0.01
3	0.63	0.03	0.62	0.05	0.82	0.10	0.48	0.04
4	0.60	0.07	0.57	0.03	0.72	0.05	0.44	0.02
5	0.53	0.07	0.54		0.67	0.06	0.42	0.03
6	0.46	0.01			0.61	0.04	0.39	0.04
7	0.45	0.04			0.57	0.12	0.35	
8	0.41				0.45			
				ITBS Form M (New Form)				
1	3.95	3.22	2.59	1.81	3.31	2.47	3.67	2.99
2	0.73	0.17	0.78	0.19	0.84	0.14	0.68	0.08
3	0.56	0.03	0.59	0.06	0.70	0.05	0.60	0.04
4	0.53	0.09	0.53	0.02	0.65	0.08	0.56	0.03
5	0.44	0.02	0.51		0.57	0.06	0.53	0.04
6	0.42	0.04			0.51	0.09	0.49	0.02
7	0.38				0.42		0.47	

Note : Reading = Reading Comprehension, Maps = Maps and Diagrams, Math = Math Problem Solving and Data Interpretation

TABLE 6
Moments for Equating ITBS Form M to Form K

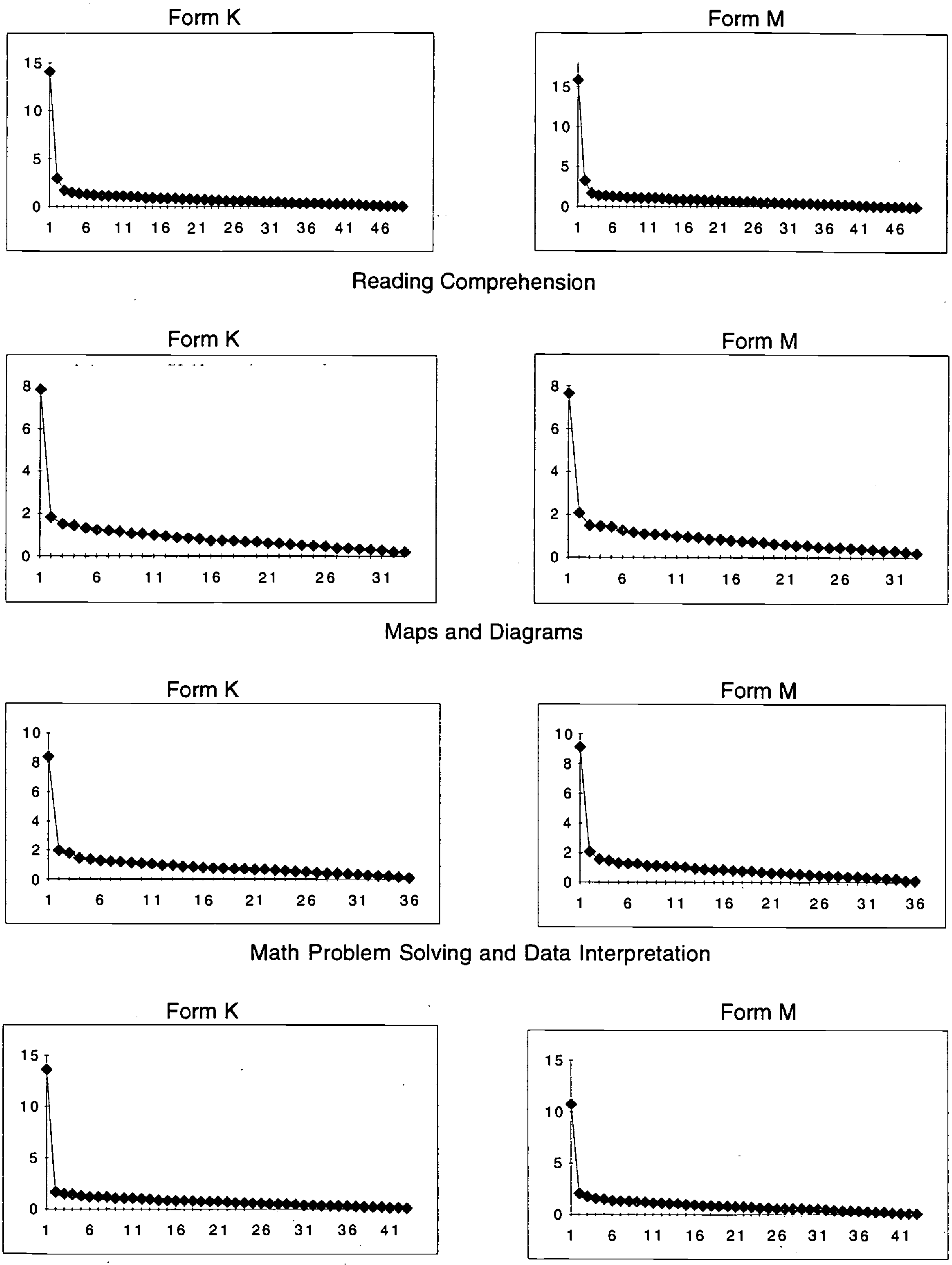
Form/ Method	Mean	Difference	S.D.	Difference	Skewness	Difference	Kurtosis	Difference
Reading Comprehension								
Form K	24.85		9.985		0.375		2.216	
Form M	25.93		10.529		0.235		2.065	
Mean	24.85	0.00	10.529	0.544	0.235	0.140	2.065	0.151
Linear	24.85	0.00	9.985	0.000	0.235	0.140	2.065	0.151
Equi_p	24.85	0.00	9.986	0.001	0.372	0.003	2.217	0.001
DIRT (T)	24.28	0.57	9.671	0.314	0.481	0.106	2.396	0.180
DIRT (O)	24.25	0.60	9.781	0.204	0.408	0.033	2.327	0.111
Nom (T)	24.82	0.03	9.841	0.144	0.337	0.038	2.322	0.106
Nom (O)	24.78	0.07	9.722	0.263	0.316	0.059	2.272	0.056
Grad (T)	25.10	0.25	9.840	0.145	0.325	0.050	2.198	0.018
Grad (O)	25.07	0.22	9.773	0.212	0.303	0.072	2.169	0.047
Maps and Diagrams								
Form K	16.28		6.337		0.383		2.306	
Form M	15.28		6.307		0.408		2.444	
Mean	16.28	0.00	6.307	0.030	0.408	0.025	2.444	0.138
Linear	16.28	0.00	6.337	0.000	0.408	0.025	2.444	0.138
Equi_p	16.27	0.01	6.326	0.011	0.376	0.007	2.291	0.015
DIRT (T)	16.03	0.25	6.022	0.315	0.576	0.193	2.582	0.276
DIRT (O)	15.98	0.30	6.085	0.252	0.487	0.104	2.542	0.236
Nom (T)	16.11	0.17	6.071	0.266	0.428	0.045	2.337	0.031
Nom (O)	16.08	0.20	6.172	0.165	0.389	0.006	2.347	0.041
Grad (T)	16.19	0.09	6.271	0.066	0.344	0.039	2.364	0.058
Grad (O)	16.21	0.07	6.273	0.064	0.345	0.038	2.363	0.057
Math Problem Solving and Data Interpretation								
Form K	16.48		6.379		0.363		2.413	
Form M	19.17		6.674		0.191		2.257	
Mean	16.48	0.00	6.674	0.295	0.191	0.172	2.257	0.156
Linear	16.48	0.00	6.379	0.000	0.191	0.172	2.257	0.156
Equi_p	16.48	0.00	6.390	0.011	0.366	0.003	2.451	0.038
DIRT (T)	16.79	0.31	6.361	0.018	0.385	0.022	2.352	0.061
DIRT (O)	16.73	0.25	6.440	0.061	0.312	0.051	2.319	0.094
Nom (T)	16.68	0.20	6.399	0.020	0.316	0.047	2.282	0.131
Nom (O)	16.63	0.15	6.466	0.087	0.297	0.066	2.254	0.159
Grad (T)	16.70	0.22	6.399	0.020	0.357	0.006	2.406	0.007
Grad (O)	16.65	0.17	6.432	0.053	0.335	0.028	2.361	0.052
Vocabulary								
Form K	24.43		8.874		0.018		2.170	
Form M	27.11		6.878		-0.244		2.623	
Mean	24.43	0.00	6.878	1.996	-0.244	0.262	2.623	0.453
Linear	24.43	0.00	8.874	0.000	-0.244	0.262	2.623	0.453
Equi_p	24.42	0.01	8.856	0.018	0.020	0.002	2.162	0.008
DIRT (T)	24.71	0.28	8.651	0.223	0.075	0.057	2.061	0.109
DIRT (O)	24.61	0.18	8.603	0.271	0.031	0.013	2.197	0.027
Nom (T)	24.35	0.08	8.589	0.285	0.008	0.010	2.294	0.124
Nom (O)	24.29	0.14	8.452	0.422	-0.022	0.040	2.339	0.169
Grad (T)	24.45	0.02	8.611	0.263	-0.180	0.198	2.452	0.282
Grad (O)	24.43	0.00	8.442	0.432	-0.181	0.199	2.483	0.313

Note: Mean = mean equating, Linear = linear equating, Equi_p = equipercentile equating, DIRT (T) = dichotomous IRT true score equating, DIRT (O) = dichotomous IRT observed score equating, Nom (T) = nominal model true score equating, Nom (O) = nominal model observed score equating, Grad (T) = graded response model true score equating, Grad (O) = graded response model observed score equating, Difference = absolute value of moment difference from form K moment using each equating method.

TABLE 7
Unweighted and Weighted Root Mean Squares for Each IRT Equating Method
Using Traditional Equating Methods as Baselines

	Unweighted Root Mean Squares			Weighted Root Mean Squares		
Method	Mean	Linear	Equipercentile	Mean	Linear	Equipercentile
Reading Comprehension Test						
DIRT (T)	1.482	1.115	0.834	1.335	1.055	0.884
Nominal (T)	0.942	0.657	0.614	0.867	0.533	0.551
Graded (T)	1.058	0.632	0.556	0.807	0.440	0.562
DIRT (O)	1.241	0.874	0.793	1.151	0.890	0.808
Nominal (O)	1.085	0.589	0.644	0.919	0.503	0.613
Graded (O)	1.071	0.518	0.572	0.832	0.398	0.572
Maps and Diagrams Test						
DIRT (T)	0.577	0.602	0.630	0.514	0.533	0.663
Nominal (T)	0.618	0.656	0.605	0.349	0.374	0.478
Graded (T)	0.376	0.393	0.351	0.172	0.190	0.380
DIRT (O)	0.426	0.462	0.547	0.426	0.447	0.592
Nominal (O)	0.444	0.484	0.471	0.287	0.309	0.435
Graded (O)	0.295	0.321	0.357	0.160	0.177	0.371
Math Problem Solving and Data Interpretation Test						
DIRT (T)	1.388	1.200	0.513	0.607	0.515	0.380
Nominal (T)	1.064	0.887	0.352	0.455	0.359	0.311
Graded (T)	1.142	0.977	0.339	0.485	0.396	0.271
DIRT (O)	0.930	0.749	0.341	0.420	0.351	0.302
Nominal (O)	0.924	0.694	0.394	0.351	0.294	0.292
Graded (O)	0.990	0.796	0.306	0.410	0.331	0.235
Vocabulary Test						
DIRT (T)	2.198	4.461	1.208	2.051	1.182	0.524
Nominal (T)	2.005	3.626	0.462	1.824	0.768	0.493
Graded (T)	2.205	3.143	0.905	1.747	0.360	0.791
DIRT (O)	1.976	3.746	0.597	1.896	0.929	0.461
Nominal (O)	1.875	3.375	0.601	1.667	0.749	0.627
Graded (O)	2.049	2.992	1.004	1.574	0.477	0.874

Note: Mean = mean equating, Linear = linear equating, Equipercentile = equipercentile equating, DIRT (T) = dichotomous IRT true score equating, DIRT (O) = dichotomous IRT observed score equating, Nominal (T) = nominal model true score equating, Nominal (O) = nominal model observed score equating, Graded (T) = graded response model true score equating, Graded (O) = graded response model observed score equating



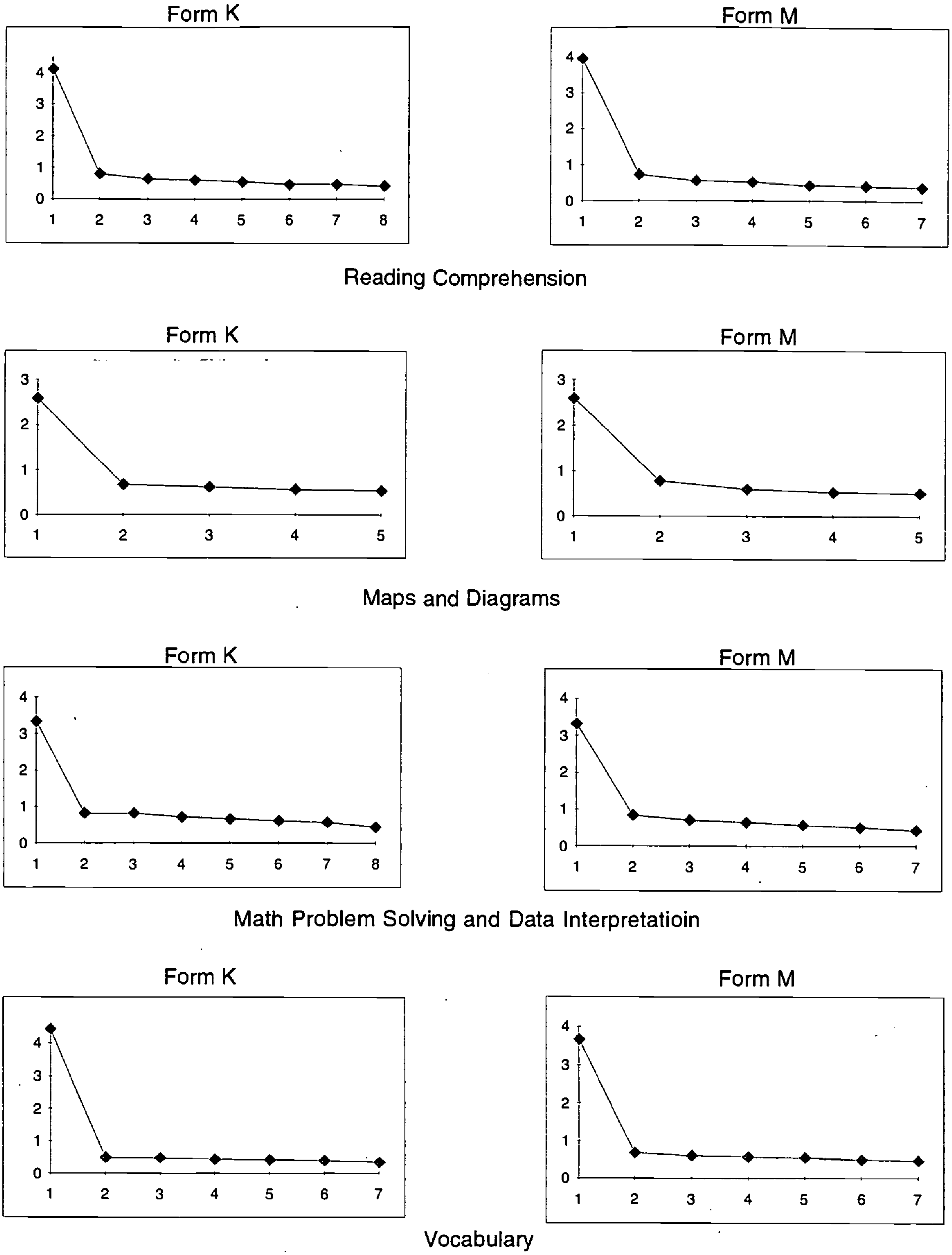
Reading Comprehension

Maps and Diagrams

Math Problem Solving and Data Interpretation

Vocabulary

FIGURE 1. Scree plots for tetrachoric correlation matrices based on individual item scores



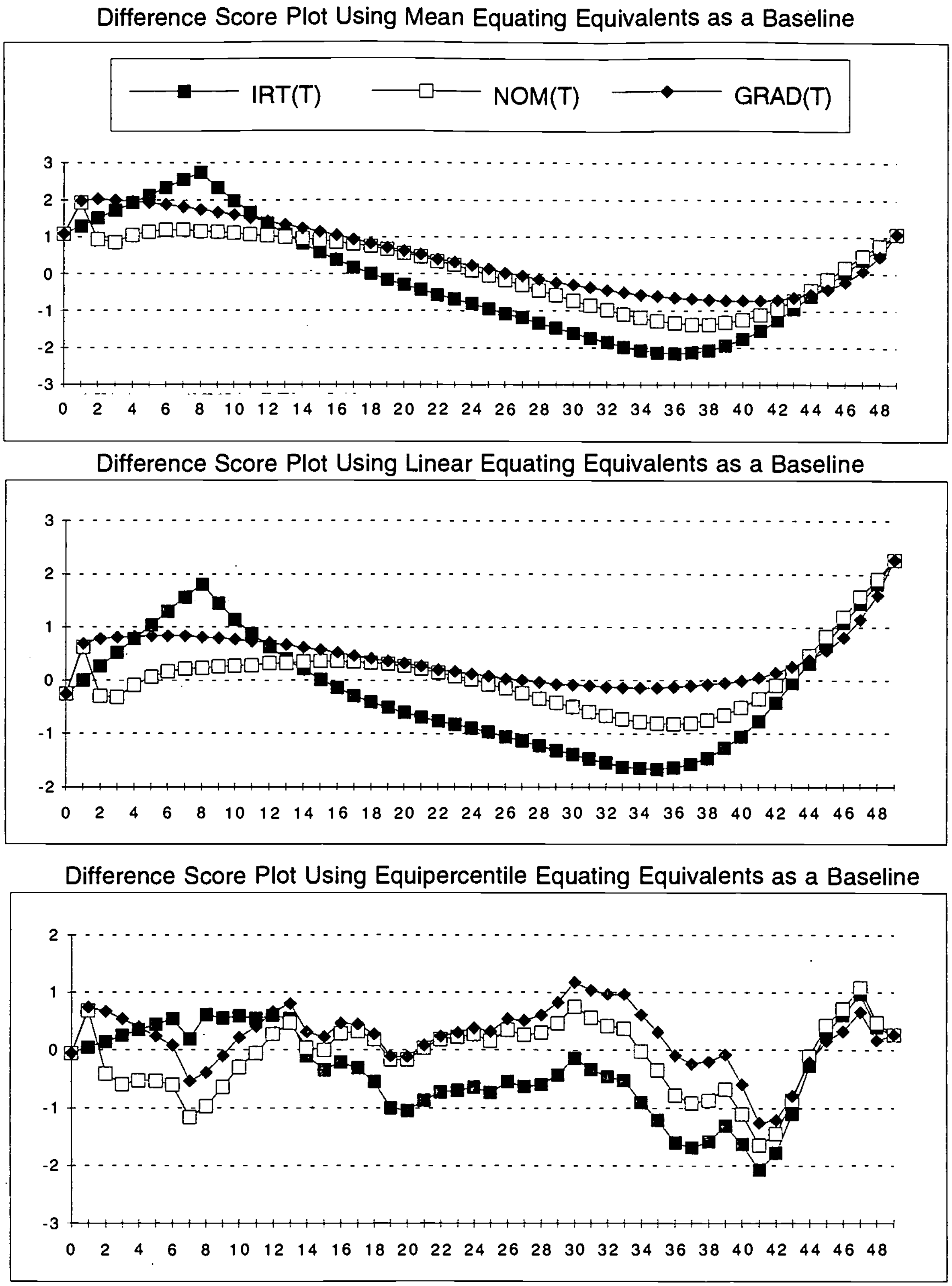
Reading Comprehension

Maps and Diagrams

Math Problem Solving and Data Interpretatioin

Vocabulary

FIGURE 2. Scree plots for product-moment correlation matrices based on passage scores

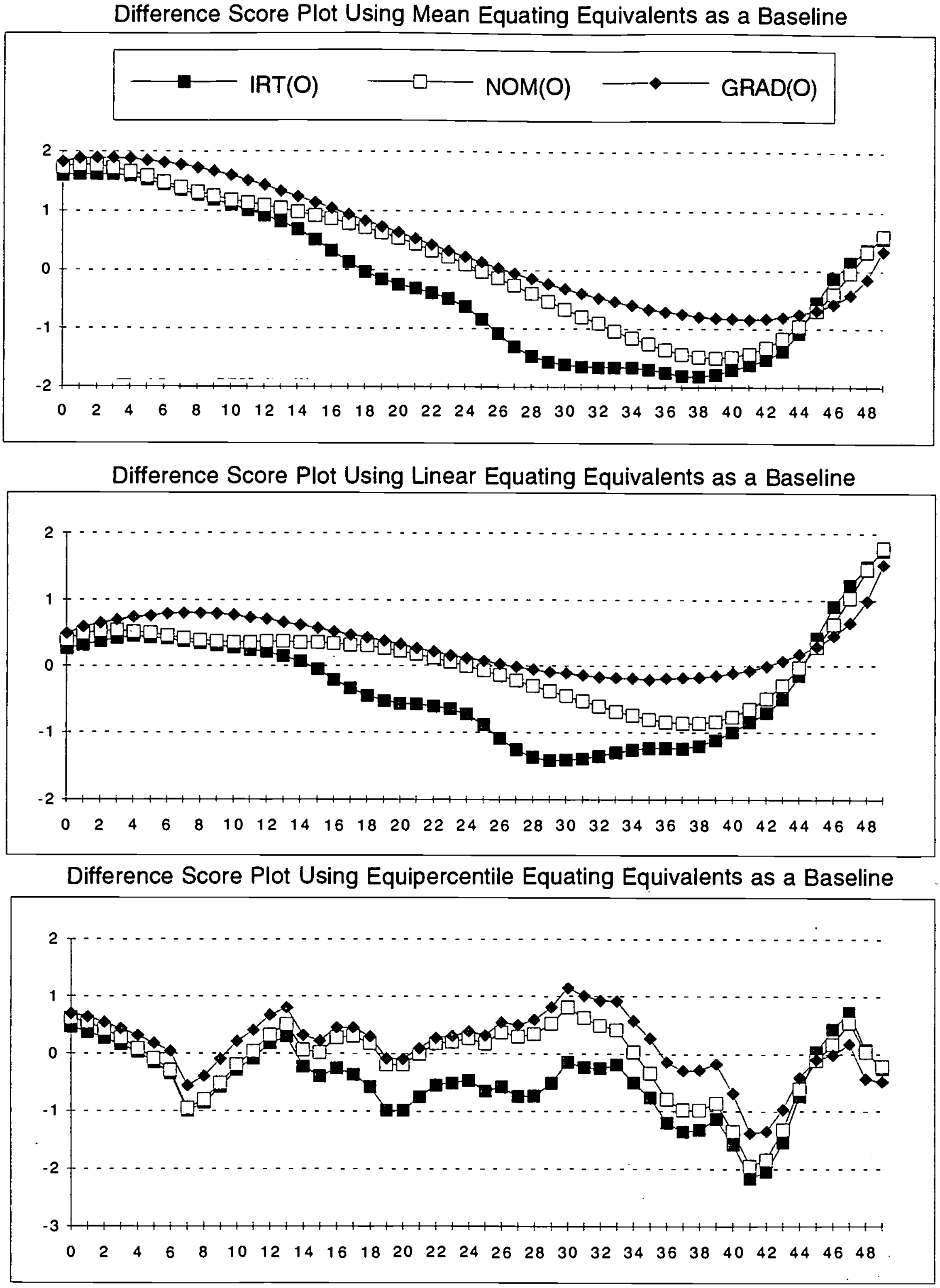


Difference Score Plot Using Mean Equating Equivalents as a Baseline

Difference Score Plot Using Linear Equating Equivalents as a Baseline

Difference Score Plot Using Equipercentile Equating Equivalents as a Baseline

FIGURE 3. Comparison of the dichotomous IRT true score equating method and the nominal model and graded response model true score equating methods using traditional equating methods as baselines for the reading comprehension tesi

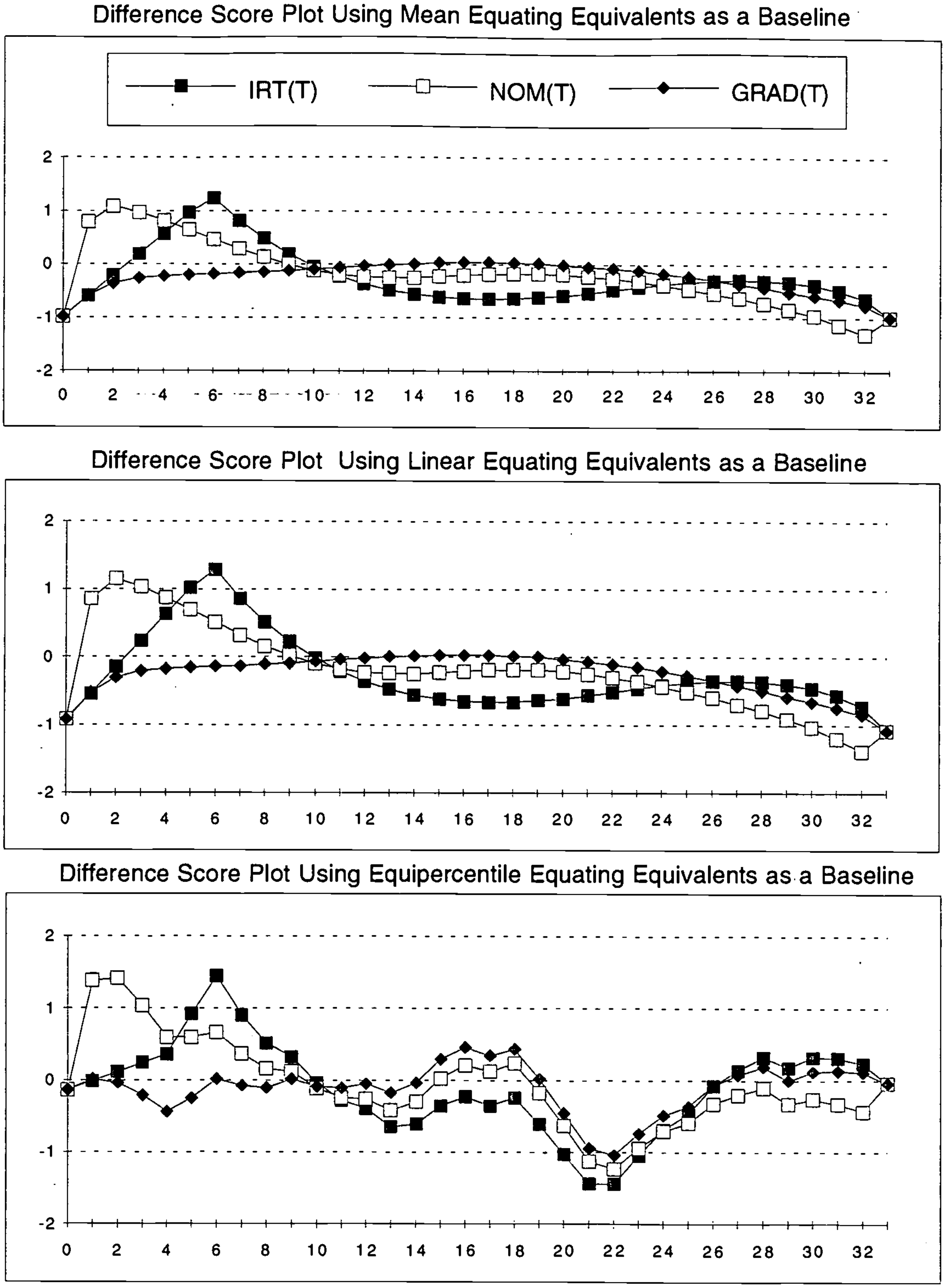


Difference Score Plot Using Mean Equating Equivalents as a Baseline

Difference Score Plot Using Linear Equating Equivalents as a Baseline

Difference Score Plot Using Equipercentile Equating Equivalents as a Baseline

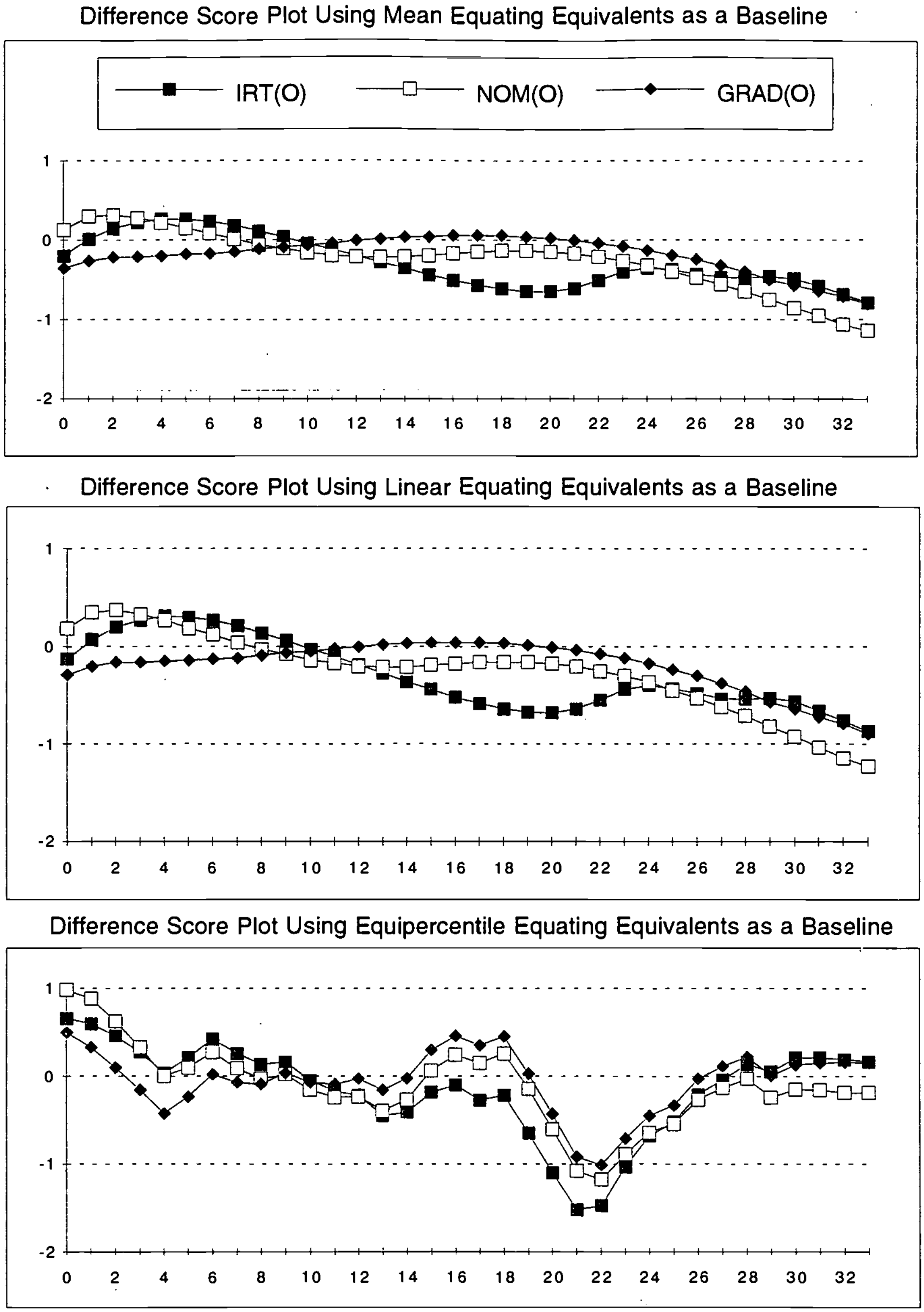
FIGURE 4. Comparison of the dichotomous IRT observed score equating method and the nominal model and graded response model observed score equating methods using traditional equating methods as baselines for the reading comprehension test


Difference Score Plot Using Mean Equating Equivalents as a Baseline

Difference Score Plot Using Linear Equating Equivalents as a Baseline

Difference Score Plot Using Equipercentile Equating Equivalents as a Baseline

FIGURE 5. Comparison of the dichotomous IRT true score equating method and the nominal model and graded response model true score equating methods using traditional equating methods as baselines for the maps and diagrams test

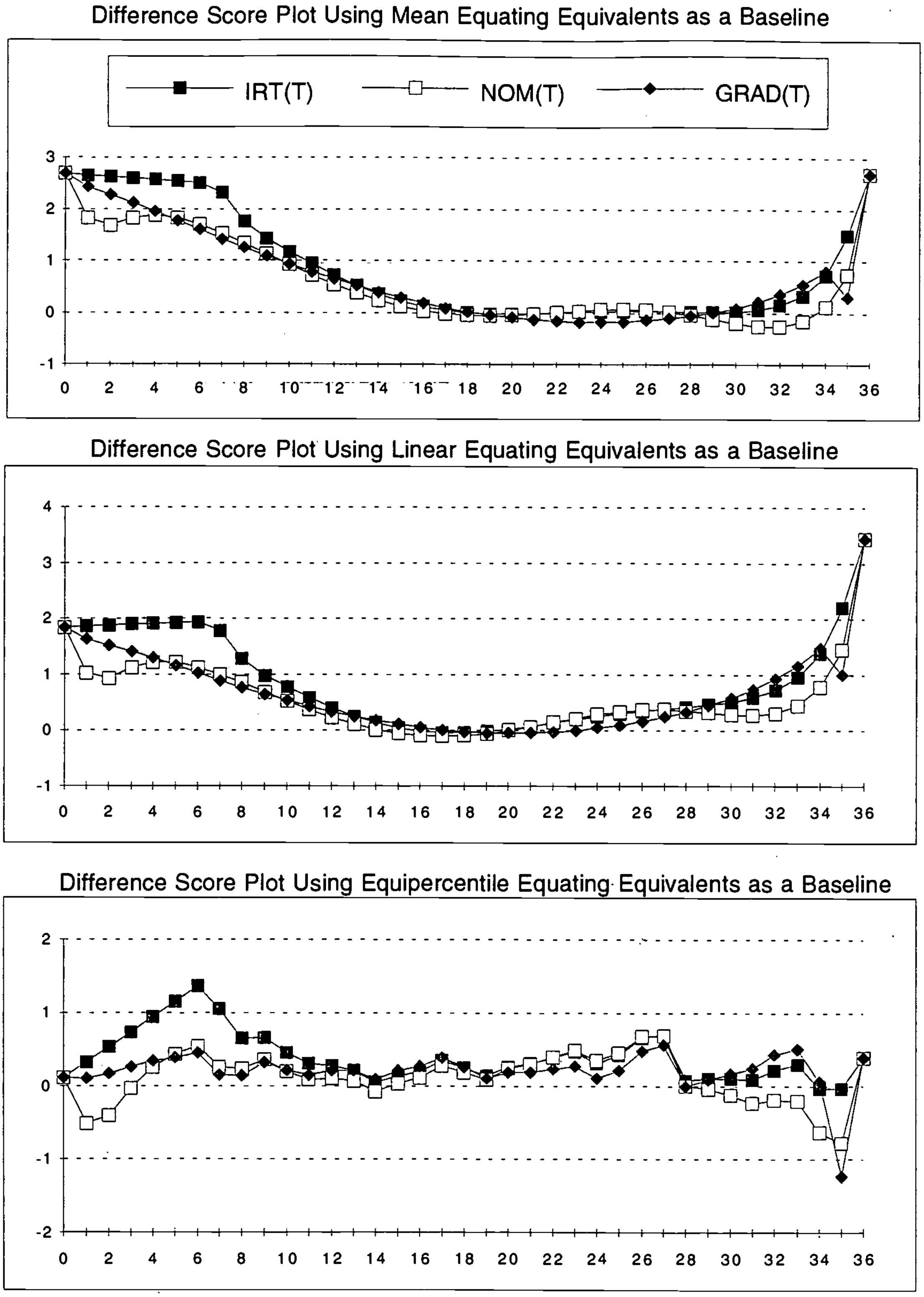


Difference Score Plot Using Mean Equating Equivalents as a Baseline

Difference Score Plot Using Linear Equating Equivalents as a Baseline

Difference Score Plot Using Equipercentile Equating Equivalents as a Baseline

FIGURE 6. Comparison of the dichotomous IRT observed score equating method and the nominal model and graded response model observed score equating methods using traditional equating methods as baselines for the maps and diagrams test

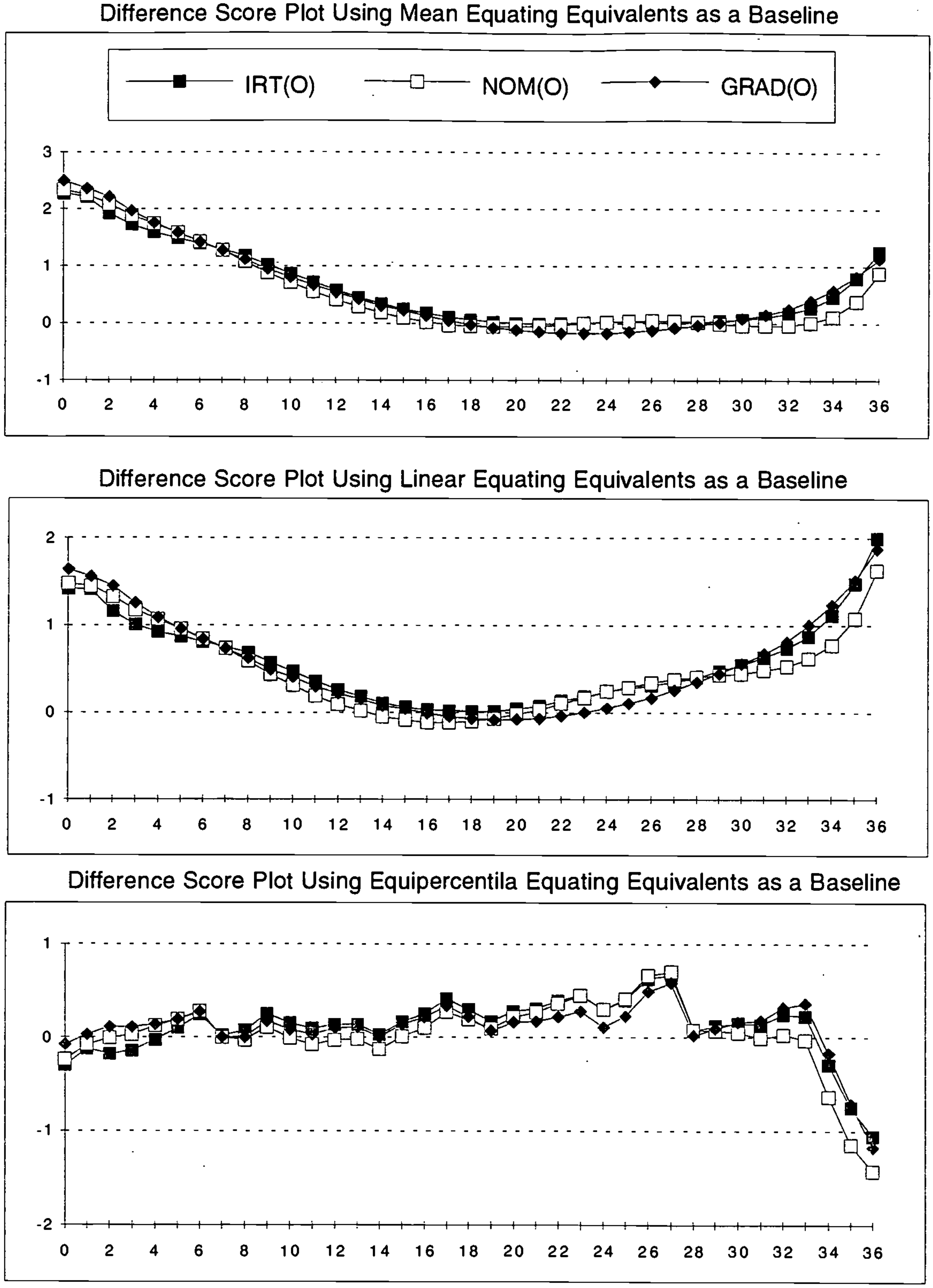


Difference Score Plot Using Mean Equating Equivalents as a Baseline

Difference Score Plot Using Linear Equating Equivalents as a Baseline

Difference Score Plot Using Equipercentile Equating Equivalents as a Baseline

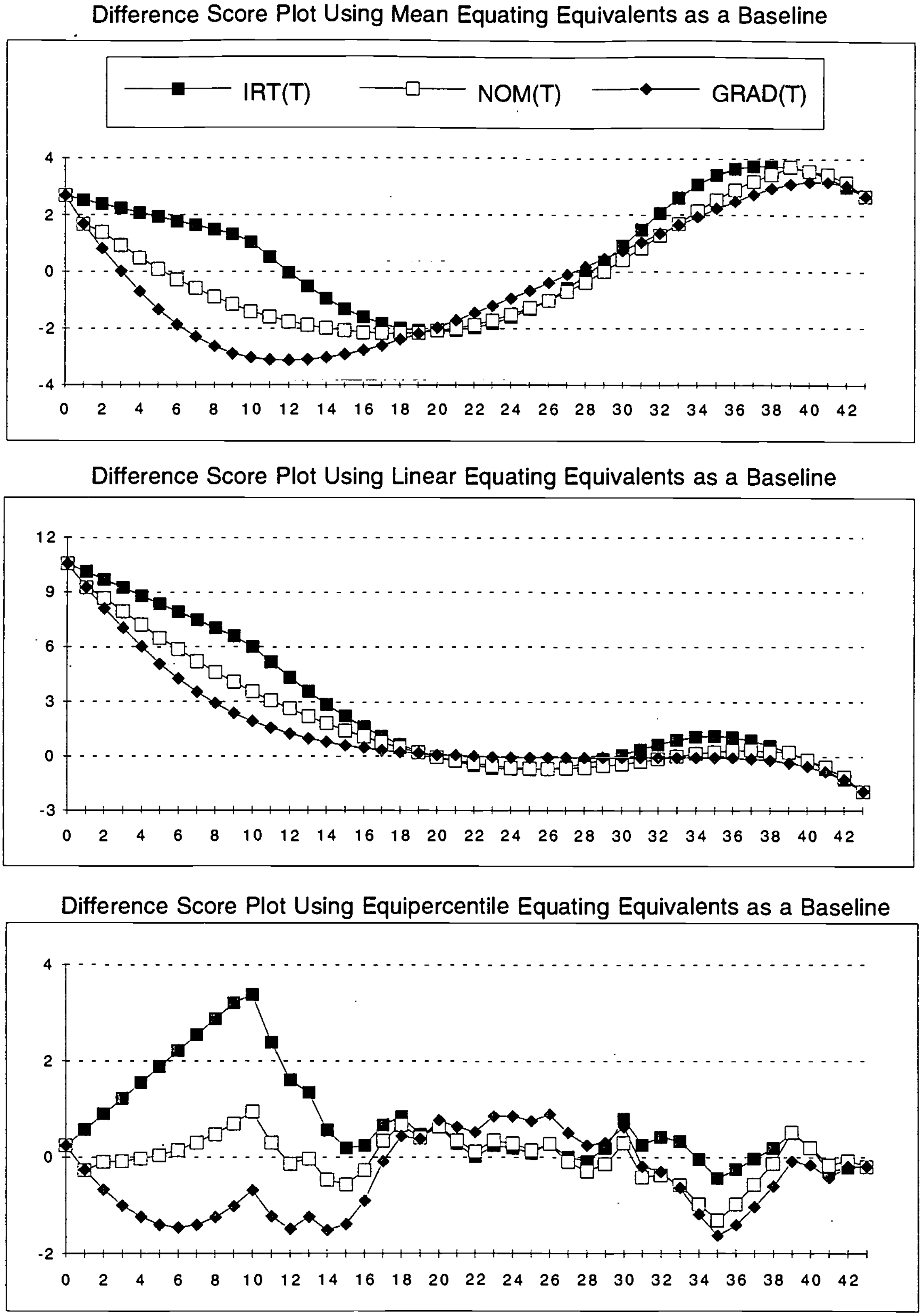
FIGURE 7. Comparison of the dichotomous IRT true score equating method and the nominal model and graded response model true score equating methods using traditional equating methods as baselines for the math problem solving and data interpretation test


Difference Score Plot Using Mean Equating Equivalents as a Baseline

Difference Score Plot Using Linear Equating Equivalents as a Baseline

Difference Score Plot Using Equipercentila Equating Equivalents as a Baseline

FIGURE 8. Comparison of the dichotomous IRT observed score equating method and the nominal model and graded response model observed score equating methods using traditional equating methods as baselines for the math problem solving and data interpretation test

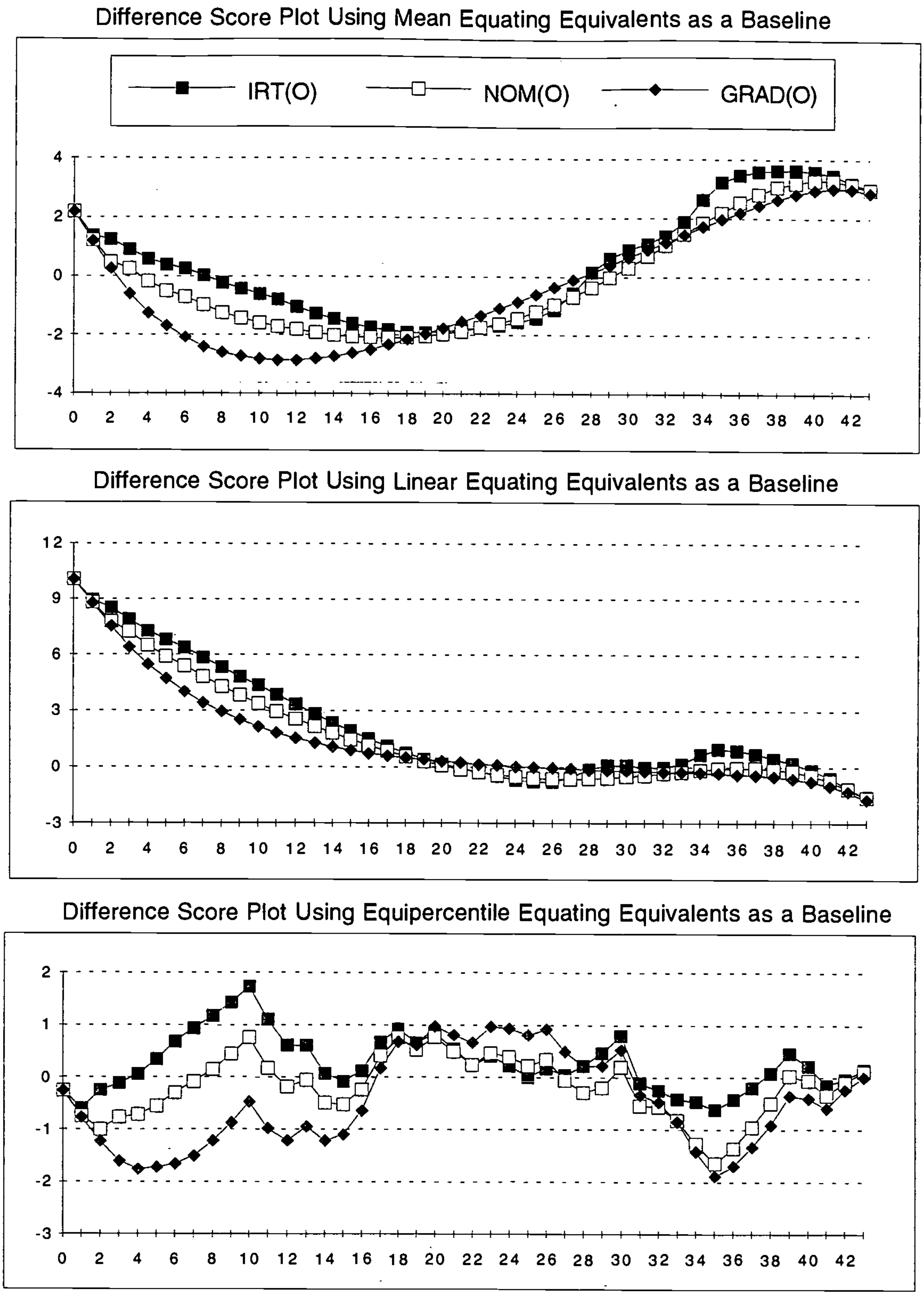


Difference Score Plot Using Mean Equating Equivalents as a Baseline

Difference Score Plot Using Linear Equating Equivalents as a Baseline

Difference Score Plot Using Equipercentile Equating Equivalents as a Baseline

FIGURE 9. Comparison of the dichotomous IRT true score equating method and the nominal model and graded response model true score equating methods using traditional equating methods as baselines for the vocabulary test


Difference Score Plot Using Mean Equating Equivalents as a Baseline

Difference Score Plot Using Linear Equating Equivalents as a Baseline

Difference Score Plot Using Equipercentile Equating Equivalents as a Baseline

FIGURE 10. Comparison of the dichotomous IRT observed score equating method and nominal the model and graded response model observed score equating methods using traditional equating methods as baselines for the vocabulary test

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM028956

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Equating Test Forms Composed of Testlets Using Dichotomous and Polytomous IRT Models

Author(s): Guemin Lee, Michael J. Kolen, David A. Frisbie, & Robert D. Ankenmann

Corporate Source: Paper Presented at the 1998 Annual Meeting of the National Council on Measurement in Education San Diego, CA	Publication Date: April, 1998

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☒	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Guemin Lee	Printed Name/Position/Title: Guemin Lee, Research Assistant	
Organization/Address: Iowa Testing Programs University of Iowa Iowa City, IA 52242	Telephone: (319)353-4721	FAX:
	E-Mail Address: gulee@blue.weeg.uiowa.edu	Date: April,23, 199

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com