MODEL SELECTION FOR EQUATING TESTLET-BASED TESTS IN THE NEAT DESIGN: AN EMPIRICAL STUDY

Wei He

Northwest Evaluation Association

Feifei Li

Educational Testing Service

Edward W. Wolfe

Xia Mao

Pearson

Paper presented at the 2012 Annual NCME Conference. Please address all questions to

Wei He wei.he@nwea.org

MODEL SELECTION FOR EQUATING TESTLET-BASED TESTS IN THE NEAT DESIGN: AN EMPIRICAL STUDY

INTRODUCTION

It is a common practice for many standardized educational tests to employ a testlet-based format in which a bundle of items share a common stimulus (e.g., a reading passage or a situated task). According to Wainer and Kiely (1987), a testlet is an aggregation of items on a single theme. A common concern that arises regarding testlet-based tests is the violation of local independence (LID) assumption.

Three methods have been proposed to address this concern. The first approach ignores the LID by fitting a standard dichotomous IRT model. The second approach involves fitting the data with a polytomous IRT model by combining the items associated with a common stimulus into one polytomous item. The last one involves modeling LID effects by building explicit models such as testlet response theory models (TRT; Bradlow, Wainer, & Wang; 1999; Wang, Bradlow, & Wainer, 2002) or the bi-factor model (Gibbons & Hedeker, 1992).

In comparison with dichotomous IRT models, the TRT model (Bradlow, Wainer, & Wang; 1999; Wang, Bradlow, & Wainer, 2002) contains an additional random effect parameter. This random effect parameter, denoted by γ, accounts for the dependency between items within the same testlet *d*. γ is allowed to vary across different testlets. The larger it is, the more item dependence the test has. A two-parameter testlet model can be expressed as:

$$P_{ij} = \frac{\exp[a_i(\theta_j - b_i - \gamma_{jd(i)})]}{1 + \exp[a_i(\theta_j - b_i - \gamma_{jd(i)})]} \quad \text{(Eq. 1)}$$

From a multidimensional modeling perspective, Li, Bolt, and Fu (2006) and Rijmen (2009) demonstrated the testlet model as a constrained version of bi-factor model (Eq. 2) in which $\gamma_{jd(i)}$ is fixed to N(0,1) and C_d is a constant for testlet *d*, which is equal to the standard deviation of $\gamma_{jd(i)}$. In Rijmen's study comparing the model equivalence among bi-factor, testlet, and second-order models, Rijmen (2009) further illustrated that the testlet model was equivalent to the second-order multidimensional IRT model for testlets, i.e., a constrained bi-factor model.

In the second-order model, items directly depend on their respective specific dimension, which in turn relies on the general dimension. Specific dimensions are conditionally independent from each other.

$$P_{ij} = \frac{\exp[a_{i1}\theta_j - a_{i1}C_d\gamma_{jd(i)} + t_i)]}{1 + \exp[a_{i1}\theta_j - a_{i1}C_d\gamma_{jd(i)} + t_i)]} \quad \text{(Eq. 2)}$$

Ignoring the testlet effects has been shown to inflate estimates of score reliability and precision (DeMars, 2006; Dresher, 2004; Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Wainer & Thissen, 1996; Yen, 1993) and yield biased item parameter estimates (Acherman, 1987; Bradlow, Wainer, & Wang, 1999; Demars, 2006; Li & Rijmen, 2009; Yen, 1993). For example, Rijmen (2009) and Li and Rijmen (2009) indicated that ignoring the testlet effects by fitting a unidimensional IRT model tended to result in a mild shrinkage of the parameter estimates. On the contrary, no scale shrinkage occurred as the result of applying bi-factor model. DeMars (2006) demonstrated that the additional parameters in the bi-factor model did not appear to decrease the accuracy of the primary trait or slope estimates.

In many large-scale testing programs, equating/linking is indispensable when different test forms are used. As discussed above that the choice of model has a significant impact on item parameter estimates in testlet-based tests, it is expected that equating results may be different given the use of different models. To the knowledge of the authors, only four relevant studies (e.g., Lee et al. (2001); Li & Cohen (2003); He et al. (2011); Tao et al. (2011)) have examined this issue. Using real data, Lee et al. (2001) explored the use of the nominal response model and the graded response model (GRM) for equating when the LID is present in the test. Their results demonstrated that the more the IRT assumptions are violated, the greater the discrepancies between equipercentile and IRT equating results. Using both IRT true- (TS) and observed-score (OS) equating and real data, Li and Cohen (2004) indicated that equating results using item parameter estimates from the TRT model were consistent with results obtained from conventional equipercentile observed score equating. Unlike polytomous IRT models, the TRT model yielded quite stable equating results across different equating methods investigated in their study. Tao et al. (2011) and He et al. (2011) respectively applied the TRT and the bi-factor models to equating tests solely composed of testlets. However, both studies used simulation; and it would be of interest to conduct a study comparing equating results from using TRT and bi-

factor models. As such, the primary purpose of this study was to apply the IRT true-score equating method to equating testlet-based tests using both TRT and bi-factor models under the non-equivalent group anchor-test (NEAT) design. In addition, the equating results from using the TRT and bi-factor models were compared with those from using conventional dichotomous IRT models.

METHODS

Data

Real data came from a state-wide reading test. Both base (Form A) and new (Form B) forms originally contain 50 items. Form A consists of 4 testlets along with 5 independent items, whereas Form B consists of 4 testlets along with 3 independent items. There is one testlet in common between Forms A and B, and this common testlet has 12 items. As this study was focused on the testlet-based test, all independent items were excluded from analysis thus reducing Form A to 45 items and Form B to 47 items. Table 1 provides an overview of the characteristics of both forms.

[Insert Table 1 about here]

Analysis Procedure

Before conducting equating, several exploratory procedures were carried out including checking total and anchor test score distributions, test dimensionality, reliability, local independence, and model-data fit. For test dimensionality, principal component analysis (PCA) was conducted via SAS. Model-data fit analysis aimed at identifying a model that could best explain the data, and the candidate models were a series of conventional IRT logistic models (i.e., 1PL, 2PL, and 3PL), TRT models with and without guessing, and the bi-factor model. Ideally, item parameter estimation is limited to one software package to prevent software differences from contaminating the final results. To this end, WinBUGS (Spiegelhalter et al., 2002) came as a natural choice, which provides a flexible and straightforward approach for calibration using the Bayesian MCMC method. For this study, the means of the Bayesian posterior distributions were used for item parameter estimates for the purpose of IRT true score equating. The priors used for the parameter estimation are presented in Table 2. The non-informative priors were used. Depending on the complexity of different candidate models, different numbers of iterations were run in WINBUGS with a single long chain ranging from

15,000 for conventional IRT models up to 70,000 for TRT models. The burn-in cycles for all models were set at 5000. Several diagnostic criteria available in WINBUGS were used to evaluate convergence including dynamic trace lines, history plots, auto-correlation lines, Gelman-Rubin convergence statistics, and quantile plots.

[Insert Table 2 about here]

An advantage that WINBUGS can provide is that the DIC that it reports can be used to evaluate model-data fit. The DIC, based on the posterior distribution of the deviance (i.e., -2(log-likelihood); denoted as D), provides a model complexity measure that can be applied to hierarchical models (Spiegelhalter et al., 2002). The DIC is defined as

$$DIC = \overline{D(\eta)} + PD$$

The first term indicates an estimated average difference between data and model, whereas the second term, called "effective number of parameters", indicates the difference between the posterior mean of the deviance and the deviance at the posterior mean of the parameters. The smaller the value of DIC is, the better the model is. According to Spiegelhalter et al. (2003), a difference of less than 5 units does not provide adequate evidence favoring one model over another.

The degree of dependency between items was evaluated using Yen's Q_3 statistic (1984) and the random effects in the TRT model, i.e., γ estimated by the WINBUGS. As mentioned above, γ can account for the dependency between items within a testlet. Studies (e.g., Wang, Bradlow, & Wainer, 2002; Wang & Wilson, 2005) suggest the magnitudes of the random effects γ .2, .9, and 1.5 representing small to large effects. To compute Q_3 for any pair of items, first, each examinee's ability estimate (denoted by θ_i) and item parameter estimates must be estimated based on the selected IRT model. Then, each examinee's expected score (E_{ij}), i.e., the probability of correct response, is computed given the selected IRT model. The difference between an examinee's observed score (O_{ij}) and the expected score, d_{ij}, can be calculated according to the following equation:

$$d_{ij=} O_{ij}\text{-}E_{ij}$$

The Q_3 value for item j and j' is the correlation of d_j and $d_{j'}$ taken over all examinees. The expected value of Q_3 is approximately -1/(n-1) if the local independence holds true, where n denotes the total number of items in a test.

Equating Design and Equating Methods

Equating Design

In this study, the NEAT design was considered. Separate calibrations followed by mean/mean linking method (Loyd & Hoover, 1980) were used to put the item parameter estimates of two forms on a common scale. Once linking constants, i.e., the slope (A) and the intercept (B), were obtained, IRT true score (TS) equating was conducted. Under the mean/mean method, the slope and the intercept constants used to put item parameter estimates on the new form to the scale of the base form were calculated in the following manner:

$$\text{A}=\bar{a}_{new}\Big/\bar{a}_{base}$$

$$\text{B}=\bar{b}_{base}-A\bar{b}_{new}$$

$\bar{a}_{base}$ and $\bar{b}_{base}$ respectively represent the means of item discrimination and item difficulty parameter estimates of the common items on the base form, whereas $\bar{a}_{new}$ and $\bar{b}_{new}$ represent the means of item discrimination and item difficulty parameter estimates of the common items on the new form. Once linking constants are worked out, the parameters on the new form can be transformed to the scale of the old form according to the following manner:

$$a^{*}_{inew}=a_{inew}/A$$

$$b^{*}_{inew}=Ab_{inew}+B$$

where * indicates a transformed value for item *i*. Note that for all models of interest in this study, the same mean/mean linking method described above was used.

IRT True Score Equating

The IRT TS equating in general involves two major steps. For the NEAT design, once item parameters for two forms are put on the same scale, the first step identifies a θ that can yield a specified true score on the new form. Using the θ value identified in the first step and the item parameter estimates on the base form, the second step looks for the corresponding true score on the base form. Kolen and Brennan (2004, p. 177-178) explain how to apply the Newton-Raphson method to find the θ value in the first step for the conventional IRT model using an iterative process.

To conduct IRT TS equating for the TRT model, one key step is to figure out how to compute the true score, which is explained in Li, Bolt, and Fu (2005). Briefly speaking, the term $\theta_j - \gamma_{jd(i)}$ in the TRT model in Equation 1 has to be reparameterized. As the result of reparameterization, the probability of answering item *i* correctly conditional on θ can be provided by the following, using the notations in Li, Bolt, and Fu (2005). Let $\xi_{jd} = \theta_j - \gamma_{jd(i)}$

$$p(y_{di} = 1 \mid \theta; \sigma_{\xi d}) = \int p(y_{di} = 1 \mid \xi_d) h(\xi_d \mid \theta; \sigma_{\xi d}) d_{\xi d}$$

h indicates the distribution of ξ_d given θ. $\sigma_{\xi d}$ are assumed known, as are the item parameters. The integral can be approximated using the Gaussian quadrature method

$$\int p(y_{di} = 1 \mid \xi_d) h(\xi_d \mid \theta; \sigma_{\xi d}) d_{\xi d} = \sum_{p=1}^{P} P(X_p) W_p$$

where X_p and W_p represent node and weight. For a test with D testlets and K items within each testlet, the true score for the whole test can be calculated as follows:

$$\tau(\theta) = \sum_{d=i}^{D} \sum_{i=1}^{k} \int p(y_{di} = 1 \mid\mid \xi_d) h(\xi_d \mid \theta; \sigma_{\xi d}) d_{\xi d} = \sum_{d=i}^{D} \sum_{i=1}^{K} \sum_{p=1}^{p} P(X_p) W_p$$

Once item parameters for two forms are put on the same scale, the true score equating for the TRT model can be conducted using the Newton-Raphson method. The mean/mean method used in the TRT model to put items on two forms on the same scale works in the same manner as that in any conventional IRT model.

With respect to the TS equating using the item parameter estimates from the bi-factor model, only the primary trait is of interest, that is, the two forms are equated only through the primary trait. As explained at the beginning of this paper, constraining the loading on the secondary dimension to be proportional to the loading on the primary dimension reduces the bi-factor model to the TRT model. This means that, in order for the t_i in Equation 2 to be comparable to the item difficulty parameter in the conventional IRT model, further transformation needs to be conducted on the parameter estimates from the bi-factor model. According to Reckase (1985), the transformation can be conducted as follows:

Let a_{i1}, a_{i2},…, a_{ik} denote factor loadings, i.e., discrimination parameters, corresponding to the *k* latent dimensions for item *i*, and d_i indicates the intercept related to an overall multidimensional difficulty for item *i*.

$$MDIFF_j = -d_j / MDISC_j$$

$MDISC_j$ can be calculated by $\sqrt{\sum_{k=1}^{k} a_{ik}^2}$. *MDIFF* can be interpreted much like the item difficulty parameter in the conventional IRT model.

Two traditional equating methods using the NEAT design were considered in this study including equipercentile and linear equating, which was carried out by CIPE (Kolen, 2004), IRT TS equating using three conventional IRT models and bi-factor model was conducted using PIE (Hanson & Zeng, 1995). IRT true score equating using TRT was completed by a program written with Matlab languages. In total, this study conducted equating eight times, described as follows. For each equating, a raw-to-raw conversion table was generated.

Equipercentile+Linear+ TS equating with *IRT models (1PL, 2PL, 3PL)* + TS *Bi-factor*+ TS *TRT models (2PL, 3PL)*

Evaluation of Equating

As in Lee et al. (2001), the results from the equipercentile and the linear equating method were used as the baselines for the reason that these methods do not assume local independence.

Three indices, defined as follows, are used to evaluate equating results: weighted bias (WBS), weighted root measure square error (WRMSE), and weighted absolute bias (WABS). In addition, the notion of difference that matters (DTM; Dorans & Feignebaum, 1994; Dorans, Holland, Thayer, & Tateneni, 2003) was adopted to evaluate the magnitude of the difference between the two conversion tables. A difference of .5 was considered significant as it resulted in the change in the reporting scores.

$$\text{WBS} = \frac{\sum_i f_i (V_i - W_i)}{\sum_i f_i}$$

$$\text{WRMSE} = \left[\sum_i f_i (V_i - W_i)^2 / \sum_i f_i \right]^{1/2}$$

$$\text{WABS} = \frac{\sum_i f_i |V_i - W_i|}{\sum_i f_i}$$

where f_i is the frequency of number-correct raw score level i, V_i is the equivalent of a number-correct score of I on the new test using an IRT method, and W_i is the equivalent of a number-correct score of i on the new test using the equipercentile method.

RESULTS

Overall Score and Anchor Item Score Distribution

Table 3 reports the descriptive statistics for total scores and anchor item scores for both forms. In general, the groups did not differ much in their ability and the forms were similar in difficulty. Both overall score and anchor item score distributions for two forms are portrayed in Figure 1 and Figure 2. For Form A, the correlation between anchor item score and total score is .808, whereas for Form B, the correlation is .814.

[Insert Table 3 about here]

[Insert Figures 1 and 2 about here]

Dimensionality

Table 4 reports eigenvalues from principal component analysis for two test forms. Clearly, each form has one dominant factor and several trivial factors, suggesting that both tests are essentially unidimensional.

[Insert Table 4 about here]

Reliability

Cronbach *a* was used to evaluate reliability for the two test forms. For the base form, Cronbach *a* is .864, whereas for the new form, Cronbach *a* is .866.

Item Dependency

Yen's Q3 statistics

Table 5 reports the mean, standard deviation of Q_3 statistics for within-testlet item pairs in base and new forms respectively. The expected Q_3 statistics, calculated by $-1/(n-1)$, were slightly different for each form. For the base form, only Testlet 1 displayed a positive Q_3 value while Testlets 2 and 4 had Q_3 values close to 0, suggesting that these testlets might exhibit item dependency. For the new form, none of the testlets display positive Q3 statistics, and Testlet 3 had Q_3 statistics close to 0, suggesting that this testlet may display more serious item dependency than others. Overall speaking, the base form may display a higher degree of violation of the item dependency than the new form.

[Insert Table 5 about here]

Random Effects γ

Table 6 reports the variance of γ for each testlet in both forms. According to the prior literature (e.g., Wang, Bradlow, & Wainer, 2002; Wang & Wilson, 2005) which suggests the magnitudes of the random effects γ .2, .9, and 1.5 representing small to large effects, both test forms did not display much serious within-item dependency. For the base form, Testlets 1 and 4 exhibit more within-item dependencies than the other two testlets; whereas for the new form, Testlet 3 exhibits the largest degree of item dependency among all four testlets. The common testlet, Testlet 2 in both forms, displayed a very similarly small degree of item dependency. It is worthwhile to note that the magnitude of random effects γ seems to depict a picture of within-item dependency similar to that by the Q_3 statistics.

[Insert Table 6 about here]

Model-Data Fit

Table 7 reports the DIC values for two forms across different models. Given the criterion that a smaller DIC value indicates a parsimonious model with better model fit, the following observations can be made: 1) among all three conventional IRT models, the IRT 3PL model fits the best; 2) among IRT 2PL, TRT 2PL, and bi-factor model, the bi-factor model appears to fit the best and fits slightly better than the TRT 2PL model; and 3) the TRT 3PL model fits slightly better than the TRT 2PL model.

[Insert Table 7 about here]

Item Parameter Estimation by the MCMC

As mentioned in the *Analysis Procedure*, a series of diagnostic tools output by the WINBUGS were use to check estimation convergence. In general, estimation convergence was satisfactory especially for the simple models such as IRT 1PL, IRT 2PL, and bi-factor models. To make sure that the estimates from the WINBUGs were comparable to those from the commercial software, they were also correlated with those from the commercial software packages including BILOG-MG and TESTFACT and publically available software SCORIGHT (Wang, Bradlow, & Wainer, 2005). Specifically, 1) the item parameter estimates for the three conventional IRT models were correlated with those from the BILOG-MG. The results indicated

very high correlation for both *a* and *b* parameters (.99 and 1 respectively), but slightly lower correlation for the *c* parameter (.81). 2) The item parameter estimates for the bi-factor model were correlated with those from TESTFACT. The results indicated the correlation coefficients of .96 for the MDIFF and .95 for the primary factor, i.e., the one used for equating. And 3) the item parameter estimates for the TRT models were correlated with those from the SCORIGHT. The results indicated the correlation coefficients ranging from .94 to .97 for all parameters except the guessing parameter, for which, the correlation coefficient was .83.

Comparisons with Different Equating Methods

Figures 3 and 4 portray the score differences conditional on raw score. The differences were computed by, at each raw score level, subtracting the equated score of the baseline equating method (equipercentile or linear) from the equated score of each equating method. The DTM bound is marked by two red lines in both figures.

Based on the criterion set for DTM, i.e., |.5|, the bi-factor and IRT 3PL models tended to yield the closest results to those by the equipercentile method with the exception that, the bi-factor model tended to produce slightly lower scores defined by the DTM at the score range 18-22, but the IRT 3PL model tended to produce slightly higher scores defined by the DTM at the high-end of the scale 35-46. The TRT 2PL yielded the second closest results to the baseline equipercentile method. The IRT 3PL—among all three conventional IRT models—provided the equating results that were closest to those by the equipercentile method. All equating methods, except the TRT 3PL and the IRT 3PL models, tended to yield lower scores than the equipercentile method, in particular, in the score range (18-29). The TRT 3PL model consistently yielded higher scores beyond that defined by the DTM than the equipercentile method except at the tails of the raw score scale. The differences in scores between the different equating methods and the equpercentile method tended to diminish toward the high-end of the scale. In general, the IRT 1PL model provided the most divergent results from those by the equipercentile methods.

[Insert Figure 3 about here]

When the linear equating results served as the baseline, the IRT 3PL—among all three conventional IRT models—provided the closest scores to the others in accordance with the criterion set for DTM, i.e., |.5|. The bi-factor model provided the second closest equating results

to those by the baseline methods. With the exception of the TRT 3PL model, all models tended to yield lower scores than those by the baseline linear method.

[Insert Figure 4 about here]

Table 8 reports the WBS, WABS, and WRMSE between the equated scores and the baseline scores. The results were not consistent when different traditional equating results served as the baselines. Among the three conventional IRT models, the IRT 2PL model provided the closest equating relationship to the equipercentile method, but the IRT 3PL model provided the closest equating relationship to the linear method. In comparison with the bi-factor model, the TRT 2PL model tended to yield closer equating relationship to the equipercentile method, though the differences in results produced by the bi-factor and the TRT 2PL models were not significantly large. However, the bi-factor model tended to yield closer equating relationship to the linear equating method than the TRT 2PL model. In general, the bi-factor and TRT 2PL models yielded the most similar equating relationship to the equipercentile baseline method than the rest of the models. And the TRT 3PL model yielded the most similar equating relationship to the linear baseline method, followed by the Bi-factor model.

[Insert Table 8 about here]

CONCLUSION AND DISCUSSION

It is accepted that testlet format can not only preserve the advantage of multiple-choice format allowing for efficient administration and objective scoring, but also provide more flexibility and efficiency in testing different aspects of cognitive activities. Literature (e.g., Demars, 2006; Rijmen, 2009; Wang, Bradlow, & Wainer, 2002) suggests that both the TRT and bi-factor models are two useful models that can be used to model data from this particular format. To the knowledge of the researchers, this study, using empirical data from a large-scale state-wide reading test, was the first study exploring equating of testlet-based tests with the TRT and the bi-factor models under the NEAT design.

In this study, specifically, several models were considered as candidate models used for IRT true score equating, and their performance were compared with those of two traditional equating methods—equipercentile and linear. Before equating was conducted, a series of

evaluation procedures were conducted including detecting item dependency level and model-data fit. For the former, both Q3 statistics and *r*—estimated by the WINBUGS—were used and results indicate a small magnitude of item dependency. Model-data fit was evaluated by the DIC values reported by the WINBUGS and the results suggest that the models that can account for the item dependency fit the data better than the conventional IRT models. When it comes to the equating, the IRT true score results in general suggest that equating using models that can account for the item dependency in general tend to yield closer equating relationship to the traditional equating methods than the conventional IRT models. This finding is in accordance with that by Lee et al. (2001), despite that the Lee et al.'s study explored the use of polytomous IRT models to equate the testlet-based tests.

It should be noted that, though using traditional equating results as baselines may sound reasonable, the generalization of the results in this study should be cautioned due to the lack of the true equating line. A simulation study which considers different levels of item dependency should be conducted in the future to investigate whether the results in this study can be replicable. At the same time, using other equating results such as the IRT observed score method and other equating designs should be explored in future studies.

REFERENCES

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for Testlets. *Psychometrika*, *64*, 153-168.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, 43(2), 145-168.

Dresher, A. R. (2004). *An empirical investigation of LID using the testlet model: A further look.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Dorans, N. J., & Feignebaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feignebaum, N. J. Deryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issued related to the introduction of the new SAT and PSAT/NMSQT* (Research Memorandum 94-10, pp.1-32). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender groups for three Advanced Placement Program exams. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (Research Report 03-27, pp. 19-36). Princeton, NJ: Educational Testing Service.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika,57,* 423–436.

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

He, W., Li, F., & Wolfe, E. (2011). *Effects of item clusters on the recovery of linking constants using test characteristic curve linking methods: Bifactor model, 2PL IRT model, and Graded Response Model*. Paper presented at the AERA annual meeting, New Orleans, LA.

Kolen, M. J. (2004). *CIPE: Common Item Program for Equating (CIPE) (version 2.0)* [computer software]. University of Iowa: Center for Advanced Studies in Measurement and Assessment (CASMA).

Lee, G., Kolen,M. J., Frisbie, D. A.,&Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement, 25,* 357–372.

Li, F., & Rijmen, F. (2009). *A vertical linking design for periodic assessment and tests that consist of situation tasks.* Paper presented at the NCME annual meeting, Saint Diego, CA.

Li, Y., & Cohen, A. (2003). *Equating tests composed of testlets: A comparison of a testlet response model and four polytomous response models*. Paper presented at the NCME annual meeting, Chicago, IL.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models of testlets. *Applied Psychological Measurement*, 30(1), 3-21.

Rijmen, F. (2009). *Three multidimensional models for testlet-based tests: Formal relations and an empirical comparison*. ETS research report: RR-09-37.

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

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D., *WinBUGS User Manual Version 1.4*, Cambridge, UK: MRC Biostatistics Unit (2002).

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., & 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.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–269). Dordrecht, Netherlands: Kluwer.

Wang, X., Bradlow, E., & Wainer, H. (2005). *Scoright (Version 3.0): A computer program for scoring tests built of testlets including a module for covariate analysis.* ETS research report: RR-04-49.

Wei, T., Cao, Y., & Lu, R. (2011). *The effect of IRT model selection on the testlet-based test equating*. Paper presented at the NCME annual meeting.

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,* 125–146.

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

APPENDIX

Table 1. *Overview of the Characteristics of the Base and the New Forms*

	Sample Size	No. Items	No. Testlet	No. Items in each Testlet
Form A	3263	45	4	13, **12**, 9, 11
Form B	3944	47	4	11, **12**, 13, 11

Table 2. *Priors Used for Model Estimation*

	a	b	c	a2	$\sigma^2_{\gamma_{d(j)}}$	$\gamma_{id(j)}$
IRT 1PL		N(0.01)				
IRT 2PL	LogN(0, 4)	N(0, .25)				
IRT 3PL	LogN (0, 4)	N(0, .25)	N(.15, 400)			
Bifactor	LogN (0, 4)	N(0, .25)		LogN (0, 4)		
TRT 2PL	LogN (0, 4)	N(0, .25)			Gamma^{-1}(1,1)	$N(0, \sigma^2_{\gamma d(j)})$
TRT 3PL	LogN (0, 4)	N(0, .25)	N(.15, 20)		Gamma^{-1}(1,1)	$N(0, \sigma^2_{\gamma d(j)})$

Note. Empty cells indicate Not Applicable

Table 3. *Descriptive Statistics for Total Scores and Anchor Item Scores for Both Forms*

	FORM A		**FORM B**	
STATS	TOTALSCORE	ANCHORSCORE	TOTALSCORE	ANCHORSCORE
MEAN	35.59	9.97	39.01	10.21
SD	6.68	1.82	6.43	1.71
MIN	4	0	8	1
MAX	45	12	47	12
MEDIAN	37	10	41	11
SKEWNESS	-1.01	-1.16	-1.34	-1.37
KURTOSIS	3.87	3.61	4.92	4.27

Table 4. *Eigenvalues from Principal Component Analysis for Two Test Forms*

	Form A	Form B
Factor1	6.905	7.468
Factor2	1.396	1.388
Factor3	1.235	1.160
Factor4	1.138	1.131
Factor5	1.103	1.123
Factor6	1.083	1.102
Factor7	1.067	1.045
Factor8	1.038	1.022
Factor9	1.026	1.001
Factor10	1.001	0.999

Table 5. *Q_3 Statistics for Within-Testlet Item Pairs*

Base Form					
Testlet	1	2	3	4	Expected
Mean	0.005	-0.009	-0.018	-0.003	-0.023
SD	0.024	0.025	0.022	0.028	
New Form					
Testlet	1	2	3	4	Expected
Mean	-0.019	-0.010	-0.004	-0.017	-0.022
SD	0.025	0.024	0.029	0.025	

Note. The expected value of Q_3 is -.023 for the base form, whereas it is -.022 for the new form.

Table 6. *Random Effects r in Each Testlet*

Base Form				
Testlet	1	2	3	4
TRT2	0.181	0.087	0.098	0.123
TRT3	0.191	0.092	0.098	0.133

New Form				
Testlet	1	2	3	4
TRT2	0.061	0.084	0.105	0.066
TRT3	0.067	0.092	0.107	0.074

Table 7. *DIC Values for Different Models*

	Base Form			**New Form**		
	IRT	TRT	Bi-factor	IRT	TRT	Bi-factor
1PL	125148	~	~	135401	~	~
2PL	123718	123446	123104	133382	133334	133021
3PL	123531	123327	~	133195	133135	~

Note. ~ indicates not applicable.

Table 8. *WBS, WABS, and WRMSE for Each Model Using Equipercentile and Linear Methods as Baselines*

	WBS		**WABS**		**WRMSE**	
	Equi%ile	Linear	Equi%ile	Linear	Equi%ile	Linear
IRT1	-0.303	-0.670	0.433	0.838	0.652	1.068
IRT2	-0.065	-0.301	0.324	0.618	0.416	0.741
IRT3	0.587	-0.221	0.607	0.423	0.631	0.491
BIFAC	0.287	-0.079	0.326	0.369	0.346	0.424
TRT2	0.136	-0.231	0.268	0.542	0.339	0.632
TRT3	0.620	0.254	0.620	0.279	0.714	0.412

Figure 1. Overall score distribution

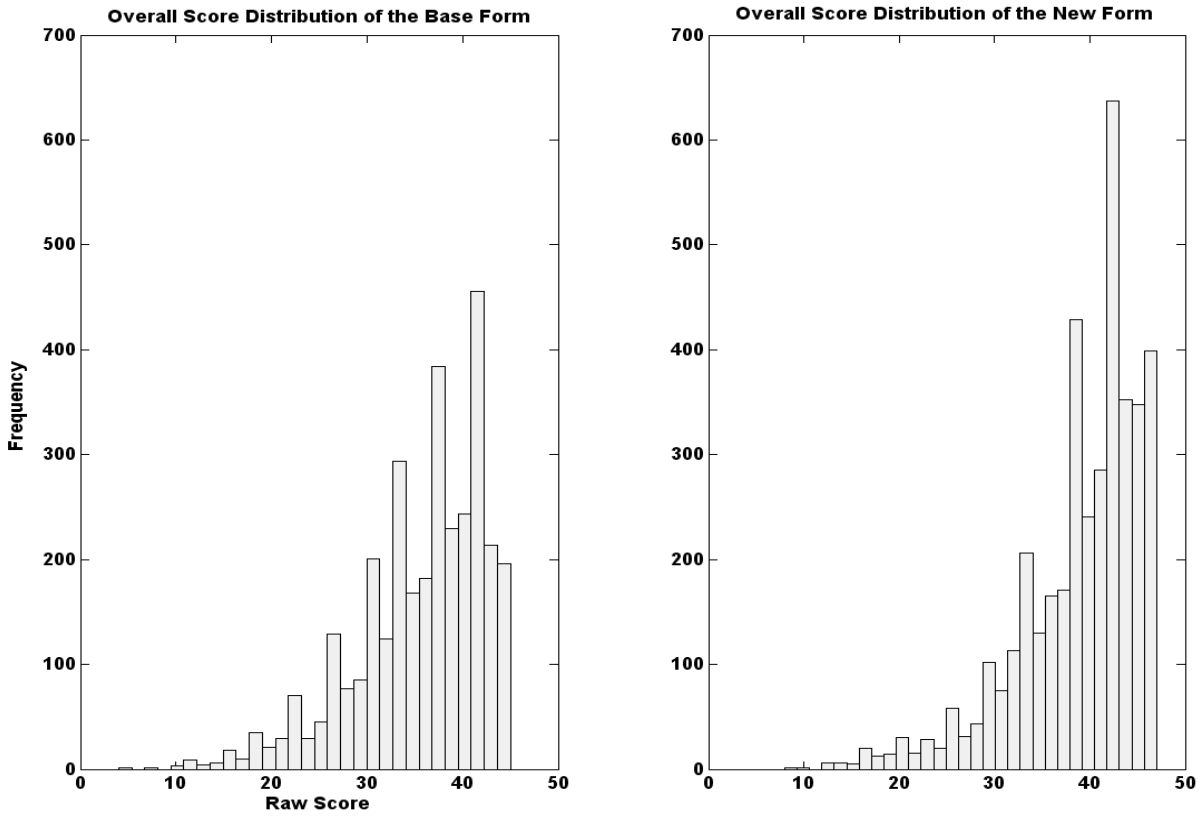

Figure 2. Anchor item score distribution

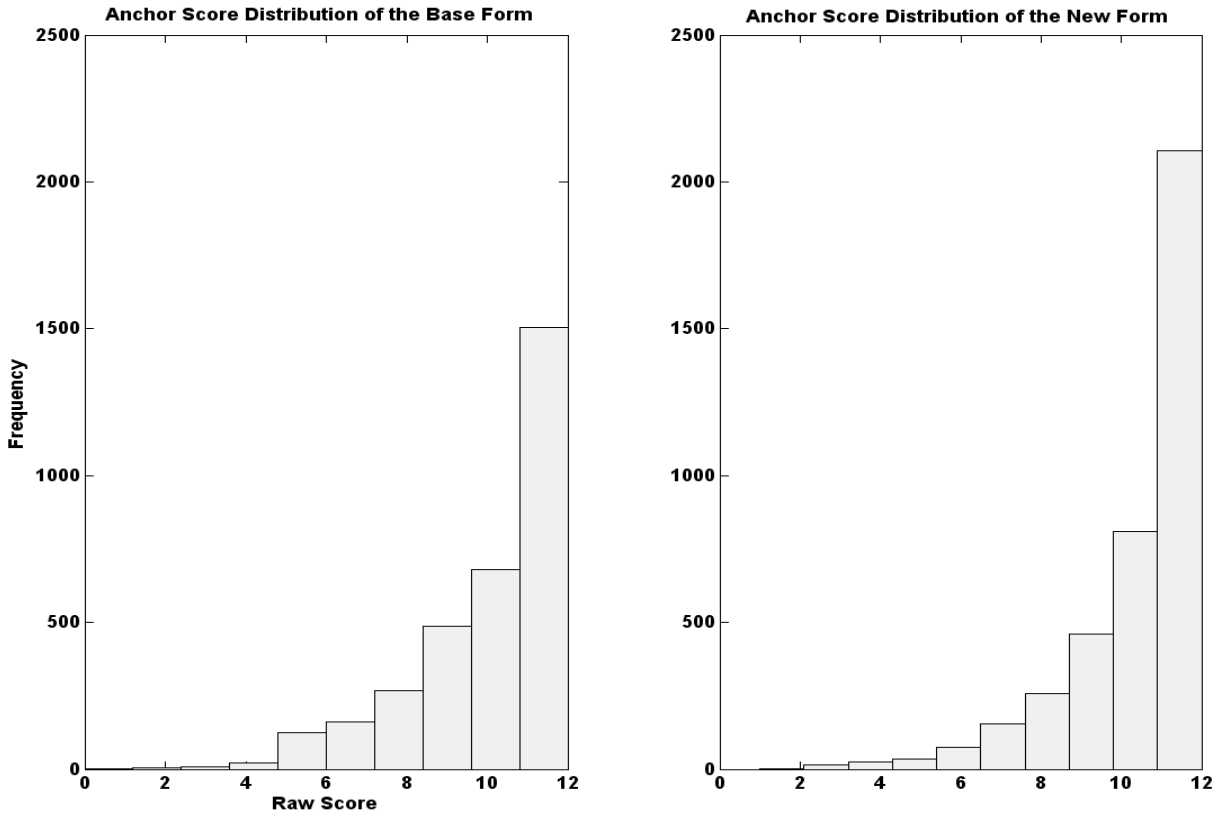

Figure 3. Difference in equating scores between IRT TS equating and equipercentile equating.

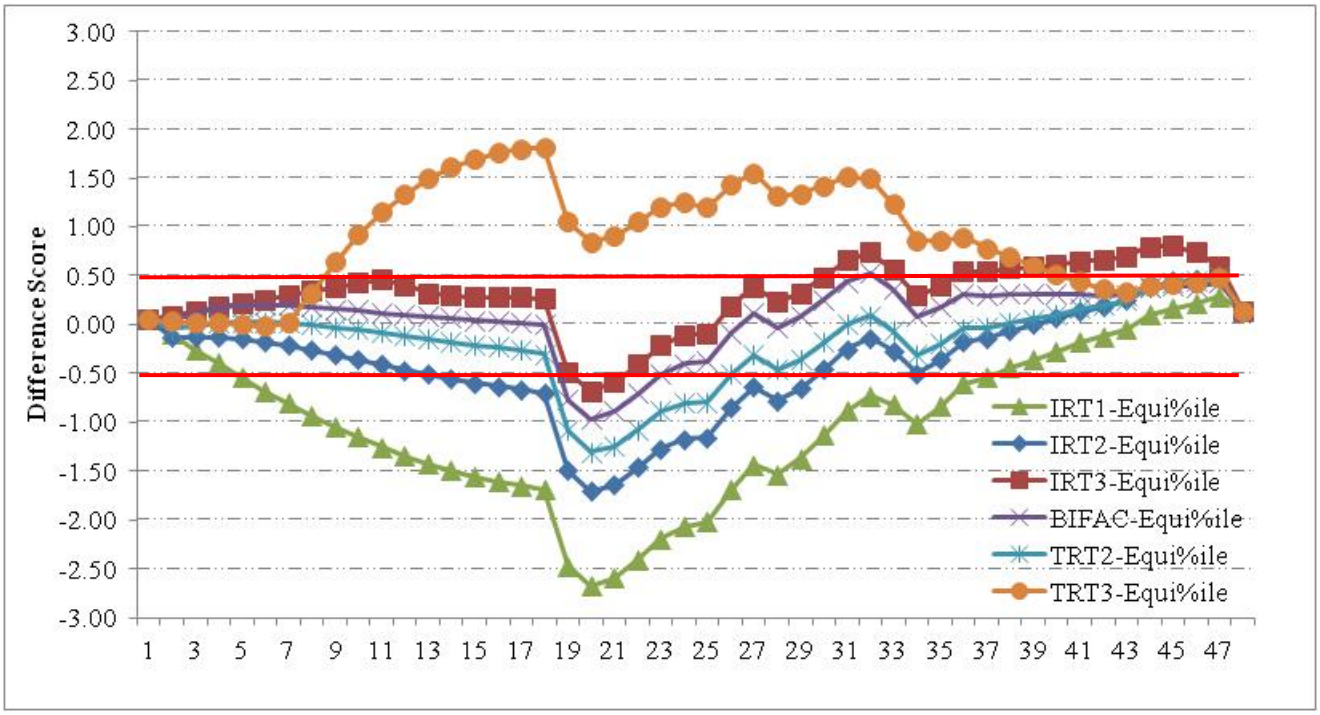

Figure 4. Difference in equating scores between IRT TS equating and linear equating.