Spaan Fellow Working Papers in Second or Foreign Language Assessment

Volume 7: 85–98

English Language Institute
University of Michigan
www.lsa.umich.edu/eli/research/spaan

Investigating Different Item Response Models in Equating the Examination for the Certificate of Proficiency in English (ECPE)

Tian Song
Michigan State University

ABSTRACT When item response models are applied in equating, the assumption of local independence is required. Polytomous item response theory (IRT) models can be considered as alternatives to dichotomous models if the assumption is violated. This study compares the performance of the dichotomous IRT model and a combination of dichotomous and polytomous IRT models in equating two forms of the Examination for the Certificate of Proficiency in English (ECPE). Traditional equating methods are used as a baseline for comparison. The results reveal that a combination of the three-parameter logistic model and the generalized partial credit model yield results similar to the traditional equating functions for the listening section, and the three-parameter logistic model performs better in the GCVR section.

In high-stakes testing programs, there is a concern that different forms might differ in difficulty, and scores on the forms are not comparable. Equating is "a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably" (Kolen & Brennan, 2004, p. 2). Therefore, test equating is a requirement for fairness to examinees.

In the literature, a wide array of equating procedures have been developed, which can be categorized into two major methods, namely traditional methods (mean, linear, and equipercentile equating), and item response theory (IRT) methods (true-score and observed-score equating). When item response models are applied in equating, the assumption of local independence is required. The local independence assumption states that after taking into account examinee ability, examinee responses to the items are statistically independent. In other words, for a given examinee, the responses to different items are not related. For example, one item does not provide clues to the correct answer to another item. However, for those items based on a common stimulus, such as reading passages or charts, local independence likely would not hold (Yen, 1993; Wainer & Thissen, 1996). In this situation, the use of dichotomous IRT models to equate tests might cause a problem. To address this problem, items associated with a common stimulus could be scored as a testlet, with summed scores of the items producing a total score for that testlet. The testlet could then be treated as a single polytomous item.

A previous study by Lee, Kolen, Frisbie, and Ankenmann (2001) has demonstrated that for tests composed of testlets only, equating based on polytomous IRT models produces results that more closely agree with the results of traditional methods than they do with dichotomous models, where the violation of the local independence assumption is severe. The present study is closely related to Lee et al.'s study, but differs in an important aspect, in that I extend the comparison of dichotomous and polytomous item response models in equating tests composed of testlets only to mixed-format tests. This study provides new evidence on the performance of different IRT models in equating tests. A mixed-format test is a test containing a mixture of different item formats (e.g., a mixture of multiple-choice and constructed response items), and is more widely used in classroom and large-scale assessments. As Baker and Kim (2004) indicated, there are many combinations of dichotomous and polytomous models that can be used to analyze data from mixed-format tests, such as a combination of the three-parameter logistic (3PL) model and graded response (GR) model, and a combination of the 3PL and generalized partial credit (GPC) model.

The primary purpose of this study is to compare equating results based on dichotomous and a combination of dichotomous and polytomous IRT models. Because traditional equating methods, such as mean, linear, and equipercentile equating, use total test scores and are not affected by the violation of local independence, they are considered as baselines for comparison. To be specific, the research question addressed is: In equating mixed-format tests, which IRT model produces the results that more closely agree with the results of traditional methods?

Competing IRT Models

Under item response theory, the interactions of a person with test items can be adequately represented by a probabilistic expression. That is, the probability of correct response to a given item is a function of both the characteristics of person and items. Over the last few decades, the use of IRT models, such as three-parameter logistic (3PL) model (Birnbaum, 1968), the generalized partial credit (GPC) model (Muraki, 1992), the graded response model (Samejima, 1969), and the nominal response model (Bock, 1972), has grown considerably in practical testing programs. In this study, the 3PL model is used for dichotomously scored items and the GPC model is for polytomously scored items.

Three-Parameter Logistic (3PL) Model

The 3PL model is the most general model for scoring dichotomous items. Under the 3PL model, the probability of examinee *i* giving a correct response for item *j* is

$$P(U_{ij}=1 \mid \theta_j, a_i, b_i, c_i) = c_i + (1-c_i)\frac{e^{Da_i(\theta_j - b_i)}}{1+e^{Da_i(\theta_j - b_i)}}$$

where U_{ij} represents the person *i*'s score on the test item *j*,

a_i is the discrimination parameter,

b_i is the difficulty parameter,

c_i is the pseudo-guessing parameter,

θ_j is the examinee's ability, and

D is the scaling constant (typically 1.7).

Generalized Partial Credit (GPC) Model

For those items based on a common passage, scores are summed for the items to produce a total score for that passage, and then the polytomous IRT model is applied to it. For example, a five dichotomous item reading passage could be treated as a polytomous item ranging from 0 to 5. In this study, the generalized partial credit model is used. The mathematical expression for the generalized partial credit model is given below

$$P(u_{ij} = k \mid \theta_j) = \frac{e^{\left[\sum_{u=1}^{k} Da_j(\theta_j - b_i + d_{iu})\right]}}{\sum_{v=1}^{m_i} e^{\left[\sum_{u=1}^{v} Da_i(\theta_j - b_j + d_{iu})\right]}},$$

where k is the score on the item,

m_i is the total number of score categories for the item,

d_{iu} is the threshold parameter for the threshold between scores u and u-1,

a_i is the overall discrimination of the item, and

b_i is the overall difficulty of the item.

Here a_i is assumed to be the same at all thresholds, but may differ across items. The threshold parameter d_{i1} is defined as 0, and d_{ik} indicates where the probability of responses changes from being greater for score category $k - 1$ to being greater for score category k. Usually the sum of the d_{ik} parameters is constrained to 0 for estimation purposes (Reckase, personal communication, February 2008).

Equating Methods

IRT Equating

Equating with item response theory is simply to put item parameter and ability estimates from two forms on a common scale. There are two major ways to develop a common scale. It can be constructed by simultaneous estimation of item parameters on a combined dataset from two forms (concurrent calibration), or an alternative way is to estimate item parameters for the two forms using two separate runs of the software, then apply linking methods (mean/mean, mean/sigma, Haebara, or Stocking and Lord method) to put them on a common scale. However, Kolen, and Brennan (2004) point out that reporting IRT ability estimates has a few disadvantages in practical testing programs. First, it is difficult to explain to examinees why the same number-correct score may receive different ability estimates. Second, examinees located at the lower and upper end of the distribution often have relatively greater amount of measurement errors. Therefore, it is better to convert estimated IRT abilities to number-correct (NC) scores and develop a relationship between NC scores on two forms.

True and observed score equating are the two methods currently available for conducting IRT equating. In true score equating, for a given ability value of theta, the number-correct true scores associated with this theta on two forms are considered to be equivalent. In IRT, the number-correct true score that is equivalent to θ_j is defined as

$$\tau(\theta_j) = \sum_i p_{ij}(\theta_j; a_i, b_i, c_i)$$

where p_{ij} is the probability of examinee *i* giving a correct answer for item *j,* and summation *j* is over items. This equation is also referred to as the test characteristics curve, which relates IRT ability to number-correct true score. Basically, there are three main steps in true scoring equating:

1. Specify a true score in one form.
2. Find the θ that corresponds to that true score.
3. Find the true score on another form that corresponds to that same θ. The true score in the last step will be considered as the equated score.

Observed score equating is conducted by estimating the frequency distributions of number-correct observed scores using item parameter and θestimates for each form, and then using conventional equipercentile equating method to approximately equate these estimated observed scores. For example, in dichotomous IRT observed score equating:

1. For one form, use the compound binomial distribution (Lord and Wingersky, 1984) to generate the distribution of observed number-correct scores for examinees with a given θ, which is denoted as $f(x|\theta)$.
2. Accumulate the observed-score distribution for examinees at each θ, and get the observed-score distribution for examinees of various abilities using

$$f(x) = \int_\theta f(x|\theta)\,\varphi(\theta)d_\theta$$

   with $\varphi(\theta)$ is the distribution of θ
3. Follow a similar procedure and get the observed-score distribution for the other form.
4. Apply equipercentile methods to equate scores from two forms.

For polytomous IRT models, the procedure is similar except that a generalization of compound binomial distribution—compound multinomial distribution—is used to model the observed number correct score distribution (Thissen, Pommerich, Billeaud, & Williams, 1995).

Traditional Equating

Two traditional equating methods used in this study are linear and equipercentile equating. In linear equating, the equating function is developed by setting the standardized deviation scores on the two forms to be equal. The equated scores deriving from this method have the same mean and standard deviation as the original scores. Equipercentile equating is to identify scores on one form that have the same percentile ranks as scores on another form. Both methods use total test scores, and are not affected by IRT assumptions. Therefore it is reasonable to consider them as baseline methods for comparing the performance of IRT models in equating.

Method

Data

The data are from the Examination for the Certificate of Proficiency in English (ECPE), which is an English as a second or foreign language test battery designed for individuals who have advanced-level language proficiency (English Language Institute, 2006, 2008). The ECPE is developed by the English Language Institute at the University of Michigan (ELI-UM), and is administered annually at approximately 125 authorized test centers in approximately 20 countries. There are four sections in each test: speaking, writing, listening, and grammar/cloze/vocabulary/reading (GCVR). The four sections are individually scored and examinees are awarded a Certificate of Proficiency based on their aggregated scores of these four sections.

Only the listening and GCVR sections are investigated in this study. All test items in these two sections are multiple-choice items. The listening section has 50 items: the first 35 items are individual items, each based on one short conversation or question; the last 15 items are based on three long dialogues, each dialogue having 5 questions. The GCVR section has 100 scored items: 30 individual items for grammar, 30 individual items for vocabulary, 20 cloze items sharing one passage, and 4 reading passages with 5 questions for each. For those items based on common reading passages or dialogues, the local independence assumption likely would be violated, and polytomous IRT models might be considered as an alternative to dichotomous models.

The two ECPE forms from year 2004–05 and 2006–07 are equated. Across the two forms there are 10 common items in the listening section, and 20 common items in the GCVR section: 10 grammar and 10 vocabulary items. Table 1 displays the descriptive statistics of raw scores on these two forms.

Table 1. Descriptive Statistics for ECPE Forms

Test	Sample Size	No. Items	Mean	SD	Skewness	Kurtosis
Form 2004–05						
Listening	33027	50	38.76	6.06	-0.643	3.209
GCVR	33027	100	68.03	11.48	-0.208	3.010
Form 2006–07						
Listening	35074	50	34.03	6.80	-0.348	2.794
GCVR	35074	100	66.79	11.66	-0.164	2.980

Analysis

Calibration. There are two choices of IRT models to analyze the dataset. If each item is considered as a unit of analysis and the test as one composed of dichotomous items only, the 3PL model is used and item parameters are estimated using BILOG-MG (Zimowski et al., 2003) with default options. If those listening, cloze, and reading items sharing a common stimulus are scored as blocks, they are treated as polytomous items for analysis. To be specific, three polytomous items were created for the listening section, four polytomous items for the cloze,[1] and four for the reading passages in the GCVR section. In this case, the test was treated as a mixed-format test composed of both dichotomous and polytomous items, and a combination of the three-parameter logistic model and the generalized partial credit model is used. Item parameters are estimated using PARSCALE (Muraki & Bock, 1991).

IRT Equating. After item parameters and θ estimates had been obtained for each form, equating was performed and the scores on Form 2006–07 were transformed to those on Form 2004–05. In this study, a random-group equating design is considered. Two groups taking the tests in 2004 and 2006 are assumed to be equivalent. It is not necessary to place item parameters of the two forms on a common scale for IRT equating. This step could be skipped, but performing it will reduce estimation errors (Hanson & Beguin, 2002). Therefore, using common items' parameter estimates from two forms, a linear transformation is estimated by Stocking and Lord's (1983) characteristic curve method. STUIRT (Kim & Kolen, 2004) is applied, and it handles both the dichotomous scored and the mixed-format test. After the item parameters and ability estimates are rescaled to a common metric, IRT true score and observed equating are conducted using the computer program POLYEQUATE (Kolen & Cui, 2004) for the mixed-format test and the computer program PIE (Hanson, Zeng & Cui, 2004) for the dichotomous scored test.

Traditional Equating. For comparison purposes, linear and equipercentile equating are conducted for each section using the computer program RAGE-RGEQUATE (Zeng, Kolen, Hanson, Cui & Chien, 2004).

Evaluation Criteria. The descriptive statistics of equated score distributions (mean, standard deviation, skewness and kurtosis) for each equating method are calculated and compared. Following Lee, Kolen, Frisbie, and Ankenmann (2001), the overall level of discrepancy between each IRT equating method and traditional equating methods can be evaluated by unweighted root mean square (URMS) and weighted root mean square (WRMS). The smaller these two indices are, the more closely the IRT equating results agree with traditional equating results. The formulas for URMS and WRMS are

$$\text{URMS}=\sqrt{\frac{1}{n}\sum_{i}(a_i-b_i)^2}\ ,$$

where a_i is equated score from IRT equating,

b_i is equated score from linear or equipercentile equating,

n is number of items, and

i represents each number correct score.

[1] 20 cloze items were collapsed into one polytomous item, but the parameter estimation did not converge. Therefore, 4 6-catergory (0 to 5) polytomous items were used here.

$$\text{WRMS}=\sqrt{\frac{1}{\sum_i f_i}\sum_i f_i\left(a_i-b_i\right)^2}\ ,$$

where f_i is the frequency distribution of the number-correct score for equated form. This index takes into account the frequency distribution of equated scores. Therefore, the scores that a large proportion of examinees receive would have greater effect on this index, and the scores that no, or a small proportion of, examinees receive would have little or no effect.

Results

Descriptive Statistics of Equated Scores

In this study, the scores on Form 2006–07 are transformed to the Form 2004–05 scale. Table 2 presents the moments of equated scores for each method and the absolute value of the difference between the equated score moments and the Form 2004–05 moments (|DIFF|). For the listening section, the equating method using a combination of 3PL and GPC models yields more similar means to those of the target form 2004–05. The mean of the equated scores using 3PL&GPC-TS (true score) and 3PL&GPC-OS (observed score) were 38.68 and 38.76, respectively. The differences from form 2004–05 means are 0.08 and 0.00, which are much smaller than those of the 3PL-TS and 3PL-OS methods (0.74 and 0.65). However, the 3PL&GPC method produces much larger differences in standard deviation (SD).

Table 2. Moments for Equating Form 2006–07 to Form 2004–05 for Linear and Equipercentile Methods and IRT Methods

Test	Mean	|DIFF|	SD	|DIFF|	Skewness	|DIFF|	Kurtosis	|DIFF|
Listening								
Form 2004–05	38.76		6.06		-0.643		3.209	
Form 2006–07	34.03		6.80		-0.348		2.794	
Linear	38.76	0.00	6.06	0.00	-0.348	0.295	2.794	0.415
Equipercentile	38.76	0.00	6.07	0.01	-0.685	0.042	3.521	0.312
3PL-TS	38.02	0.74	6.06	0.00	-0.788	0.145	3.884	0.675
3PL-OS	38.11	0.65	6.02	0.04	-0.646	0.003	3.418	0.209
3PL&GPC-TS	38.68	0.08	5.82	0.24	-0.784	0.141	3.760	0.551
3PL&GPC-OS	38.76	0.00	5.86	0.20	-0.640	0.003	3.224	0.015
GCVR								
Form 2004–05	68.03		11.48		-0.208		3.010	
Form 2006–07	66.79		11.66		-0.164		2.980	
Linear	68.03	0.00	11.48	0.00	-0.164	0.044	2.980	0.030
Equipercentile	68.03	0.00	11.48	0.00	-0.217	0.009	3.081	0.071
3PL-TS	67.95	0.08	11.51	0.03	-0.153	0.055	3.092	0.082
3PL-OS	67.95	0.08	11.47	0.01	-0.154	0.054	3.064	0.054
3PL&GPC-TS	67.78	0.25	11.99	0.51	-0.186	0.022	2.756	0.254
3PL&GPC-OS	67.78	0.25	11.84	0.36	-0.179	0.029	2.738	0.272

For the GCVR section, the 3PL equating method provides more similar moments to those of the target form than the 3PL&GPC method. In terms of 3PL true scoring equating, the differences in mean, SD and kurtosis from the target are 0.08, 0.03, and 0.082 respectively, which are much smaller than those of the 3PL&GPC method. This was also true for the observed scoring equating. However, the 3PL&GPC equating method produces more similar skewness values.

Equating Conditional on NC Scores

Conditional on each number correct (NC) score, the differences between the equated score of IRT equating methods and the equated score of the baseline equating methods (linear and equipercentile equating) are calculated. The smaller the absolute value of the difference is, the more similar the equating function is to the baseline equating function. Figures 1 and 2 display the difference scores for the listening and GCVR sections, respectively, with the difference score on the vertical axis and NC score on the horizontal axis.

Listening Section. In Figure 1, in terms of true score equating, the 3PL&GPC equating function was more similar to the linear equating function than the 3PL method in the score range 15–25. For scores above 25, both the 3PL&GPC and 3PL methods produce equivalents similar to those of linear equating. For scores below 15, the differences between both methods and linear equating are very large (the differences are around -8 for the scores below 10). The cause might be that few examinees scored below 15, which yielded a large amount of equating errors. The patterns are similar for observed score equating.

In Figure 2, in terms of true score equating, the 3PL&GPC equating function is more similar to the baseline equipercentile equating function than the 3PL method for most scores between 10 and 40. For scores above 40, the equated score of the two methods are similar to those from equipercentile equating. Moreover, observed score equating performs better than true score equating, which provides more similar equivalents to those of equipercentile equating.

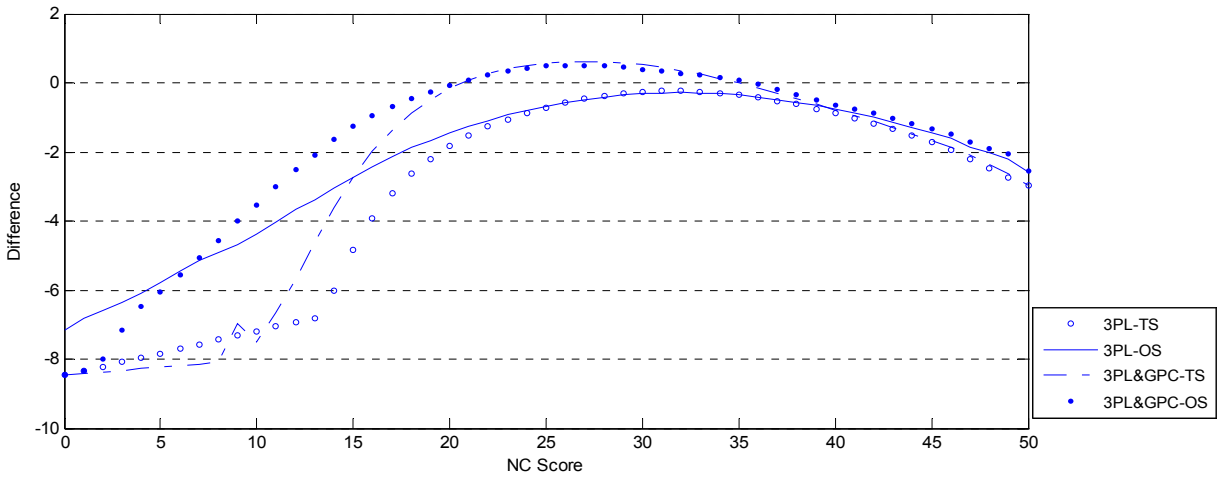

Figure 1. Comparsion of IRT Models using Linear Equating as Baseline for Listening Section

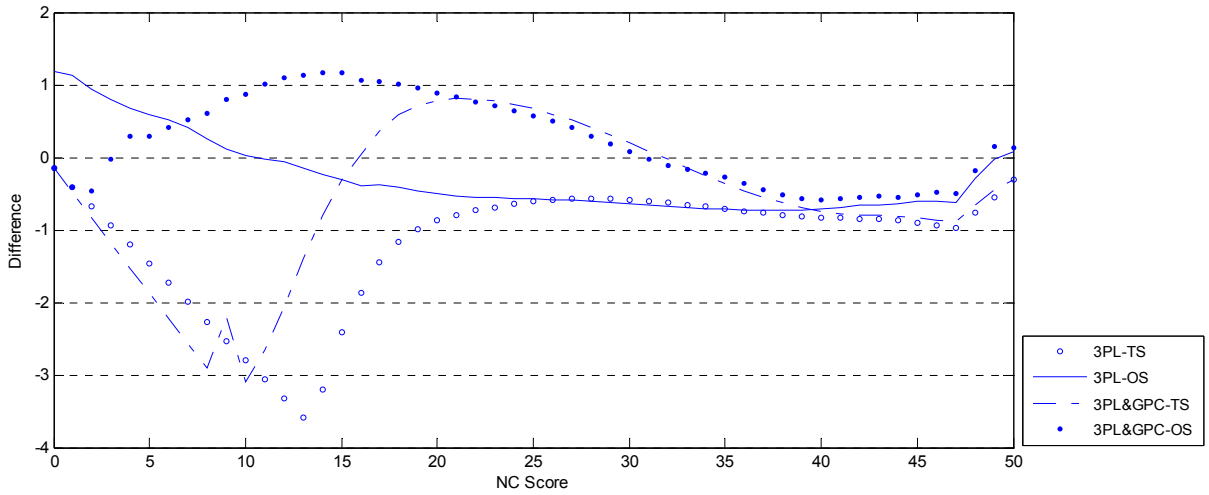

Figure 2. Comparsion of IRT Models using Equipercentile Equating as Baseline for Listening Section

GCVR Section. In both figures 3 and 4, the true score and observed score equating provides similar equating relationships except that the 3PL&GPC-TS performes differently below the score 30. The number of examinees who scored below 30 was very small; therefore no reliable pattern would be expected due to the small sample size and large equating errors. In terms of IRT models, the 3PL equating method yields more similar equivalents to those of linear and equipercentile equating than does the 3PL&GPC method for most NC scores.

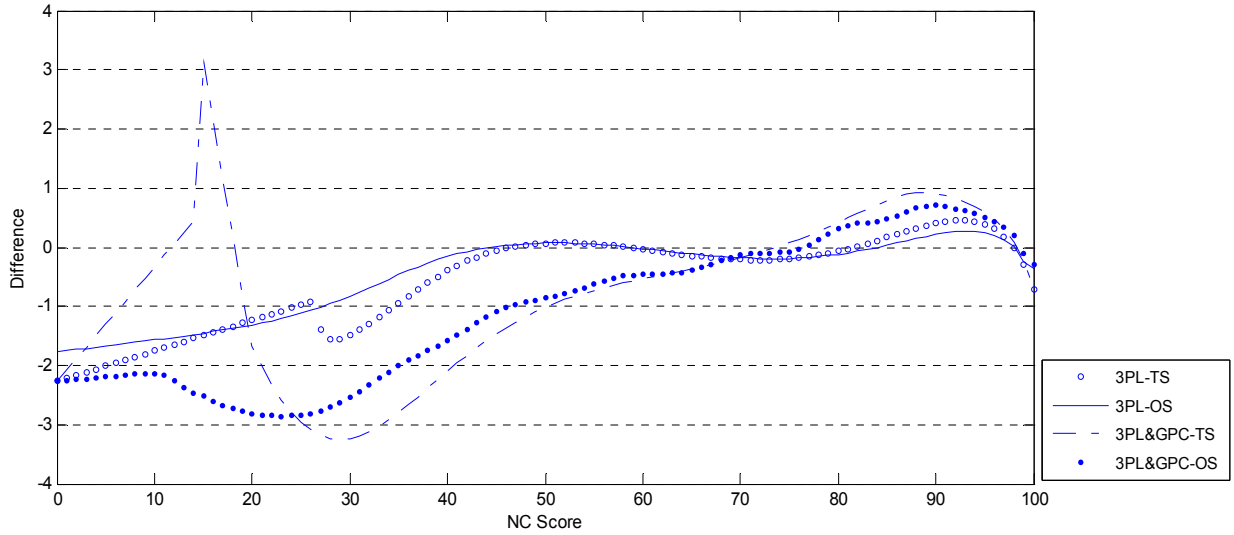

Figure 3. Comparsion of IRT Models using Linear Equating as Baseline for GCVR Section

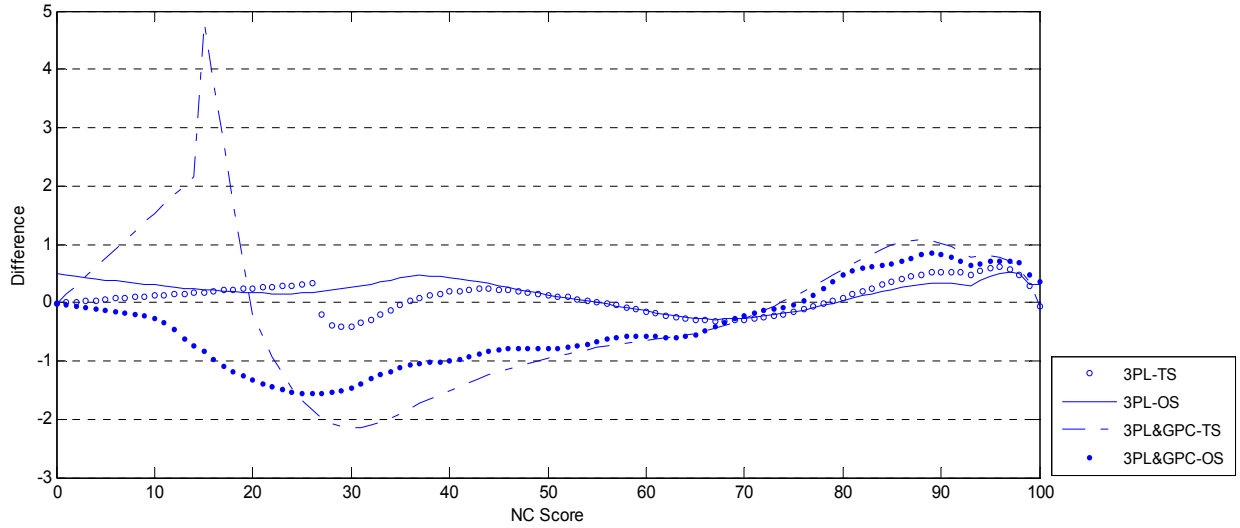

Figure 4. Comparsion of IRT Models using Equipercentile Equating as Baseline for GCVR Section

The different findings for the listening and GCVR sections might be due to the different degree of violation of the local independence assumption. If the assumption is severely violated, the use of polytomous models would eliminate the effect of dependence among items and likely to improve the equating. In this study, it seems that the IRT assumption is violated less with the GCVR section than with the listening section.

Weighted Root Mean Square and Unweighted Root Mean Square

Table 3 shows the Weighted Root Mean Square (WRMS) and Unweighted Root Mean Square (URMS) between equated scores and the baselines. These two indices represent the overall level of discrepancy between each IRT equating method and the traditional equating methods.

Table 3. WRMS and URMS for each IRT Equating Method using Linear and Equipercentile Methods as Baselines

	WRMS		URMS	
Test	Linear	Equipercentile	Linear	Equipercentile
Listening				
3PL-TS	0.985	0.761	4.408	1.452
3PL-OS	0.795	0.659	3.119	0.599
3PL&GPC-TS	0.798	0.570	4.220	1.181
3PL&GPC-OS	0.615	0.461	3.146	0.639
GCVR				
3PL-TS	0.185	0.236	0.970	0.269
3PL-OS	0.140	0.210	0.817	0.287
3PL&GPC-TS	0.662	0.669	1.545	1.314
3PL&GPC-OS	0.528	0.542	1.607	0.830

For the listening section, in terms of WRMS, 3PL&GPC equating methods provide more similar equating relationships to the baseline methods than 3PL methods. For example, using equipercentile equating as the baseline, the WRMS of 3PL&GPC-TS and 3PL&GPC-OS are 0.570 and 0.461, respectively, which are smaller than those of 3PL methods (0.761 and 0.659, respectively). However, using URMS as the criterion, two models perform similarly and observed score equating yields more consistent results with traditional equating methods than with true score equating.

For the GCVR section, 3PL equating methods yield smaller WRMS and URMS than 3PL&GPC methods. This indicates that 3PL provides more similar equating relationships to the baseline methods. This is true for either true score or observed score equating.

Summary and Discussion

As a large-scale certification test with high stakes, the ECPE needs to ensure fairness and consistency in each testing situation. The problem of comparability among test scores using different test forms must be addressed. When equating is conducted under item response theory, failing to taking into account the effect of local dependence among items might distort the equated scores, and disadvantage individual test takers.

In this study, two ECPE forms were equated using different IRT models and compared to traditional equating methods. The results reveal that a combination of 3PL and GPC models performed better than the 3PL model for the listening section, especially for low and medium scores (ranging from 15 to 25). However, for the GCVR section, the 3PL model yielded a more similar equating function to the traditional equating function. The dissimilarity between the two sections might be due to the different degree of violation of the local independence assumption.

The choices of dichotomous and polytomous models in this study were restricted. Only the three-parameter logistic model and the generalized partial credit model were investigated, thus the results from this study may not generalize to other models. More IRT models should be studied in the future, such as the graded response model and the nominal model.

In addition, only real data were analyzed in this study and the results might be limited to this data. Further research using simulation techniques needs to be pursued. Using simulation, different factors could be manipulated, such as the percentage of polytomous items in mixed-format tests, the length of tests, and the choice of IRT models. Therefore, the effects of different factors on equating relationships in mixed-format tests could be better evaluated.

Acknowledgments

The author thanks the University of Michigan English Language Institute (ELI-UM) for providing funding for this research, and also thanks Dr. Jeffrey Johnson for editing and reviewing.

References

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models and their use in inferring examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

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.

English Language Institute, University of Michigan. (2006). *Examination for the Certificate of Proficiency in English 2004–05 annual report*. Ann Arbor: English Language Institute, University of Michigan.

English Language Institute, University of Michigan. (2008). *Examination for the Certificate of Proficiency in English 2006–07 annual report*. Ann Arbor: English Language Institute, University of Michigan.

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3–24.

Hanson, B. A., Zeng, L., & Cui, Z. (2004). *PIE* [Computer Software]. Iowa City: University of Iowa.

Kim, S., & Kolen, M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models*. Iowa City: Iowa Testing Programs, The University of Iowa.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Kolen, M. J., & Cui, Z. (2004). *POLYEQUATE* [Computer Software]. Iowa City: University of Iowa.

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*(4)*,* 357–372.

Lord, F. M. (1982). Standard error of an equating by item response theory. *Applied Psychological Measurement*, *6*(4), 463–472.

Lord, F. M., & Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score "equating". *Applied Psychological Measurement*, *8*(4)*,* 452–461.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2)*,* 159–176.

Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter scaling of rating data* [Computer software]. Chicago: Scientific Software International, Inc.

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

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*(1), 39–49.

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*, 22–29.

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

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.

Zeng, L., Kolen, M. J., Hanson, B. A., Cui, Z., & Chien, Y. (2004). *RAGE-RGEQUATE* [Computer software]. Iowa City: University of Iowa.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3.0* [Computer software and manual]. Chicago: Scientific Software International, Inc.