*Center for Advanced Studies in*
*Measurement and Assessment*

*CASMA Research Report*

*Number 26*

# Assessing IRT Model-Data Fit for Mixed Format Tests[*]

*Kyong Hee Chon*
*Won-Chan Lee*
*Timothy N. Ansley*[†]

November 2007

[*] The authors are grateful to Robert Brennan for his valuable comments on the paper. A previous version of this paper was presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, April, 2007.

[†] Kyong Hee Chon is Research Assistant in Iowa Testing Programs (email: kyonghee-chon@uiowa.edu). Won-Chan Lee is Associate Director, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: won-chan-lee@uiowa.edu). Timothy N. Ansley is Associate Professor, 320 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: timothy-ansley@uiowa.edu).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

# Contents

# List of Tables

# List of Figures

# Abstract

This study examined various model combinations and calibration procedures for mixed format tests under different item response theory (IRT) models and calibration methods. Using real data sets that consist of both dichotomous and polytomous items, nine possibly applicable IRT model mixtures and two calibration procedures were compared based on traditional and alternative goodness-of-fit statistics. Three dichotomous models and three polytomous models were combined to analyze mixed format test using both simultaneous and separate calibration methods. To assess goodness of fit, The PARSCALE's $G^2$ was used. In addition, two fit statistics proposed by Orlando and Thissen (2000) were extended to more general forms to enable the evaluation of fit for mixed format tests. The results of this study indicated that the three parameter logistic model combined with the generalized partial credit model among various IRT model combinations led to the best fit to the given data sets, while the one parameter logistic model had the largest number of misfitting items. In a comparison of three fit statistics, some inconsistencies were found between traditional and new indices for assessing the fit of IRT models to data. This study found that the new indices indicated considerably better model fit than the traditional indices.

# 1 Introduction

With the increasing use of tests that contain both multiple-choice and constructed-response items, a great deal of research has been devoted to combining scores from a mixture of different item formats – referred to in this study as a mixed format test (e.g., Traub, 1993; Wainer & Thissen, 1993; Ercikan et al., 1998; Sykes & Yen, 2000). Under the item response theory (IRT) framework, an important issue in the calibration of mixed format data is whether the different item types (i.e., multiple-choice and constructed-response items) measure the same construct. It has been reported that some large-scale operational mixed format tests (e.g., Bennett, Rock, & Wang, 1991; Thissen, Wainer, & Wang, 1994) are nearly unidimensional with respect to the constructs that are measured, but others (e.g., Ercikan et al., 1998; Traub, 1993) are not (Kim, 2004). In either case, theoretical models and estimation programs are available for calibrating items in a mixed format test. Even within unidimensional IRT models, there can be a variety of model combinations because the use of multiple item formats typically involves more than one IRT model.

Calibrating two different item types raises an issue of choosing a calibration procedure, namely, the two types of item responses can be calibrated either simultaneously or separately. Simultaneous calibration can be advantageous in that the total number of items remains the same in estimation. By contrast, the number of items is necessarily decreased in separate calibration since the full-length test is divided into two shorter parts for the two item types. Lord (1980) argued that test length in combination with sample size affects the quality of parameter estimates. Hambleton and Cook (1983) suggested that at least 200 examinees and 20 items should be used for stable results of parameter estimates when scaling multiple-choice items using IRT. In this sense, maintaining the total number of items can enhance the estimation accuracy, particularly with a small sample size. Another benefit of simultaneous calibration is that it requires only a single run of an IRT estimation program. Simultaneous calibration may consist of either a mixture of different models or a single model. Alternatively, the two item types can be calibrated separately. Some measurement settings might better match this approach. For example, the two item types are often administered on two different occasions to the same group of examinees. Item responses from the two separate administrations can then be combined together for scaling or scoring purposes. In this case, separate calibrations of the two item types may be preferred.

When various models and calibration procedures are available for a particular data set, one natural question that arises is which one to choose. One way to assess the appropriateness of the chosen IRT model(s) and calibration procedure is to conduct an analysis of model-data fit. Model-data fit is regarded as a useful checking tool in model selection. As noted by many researchers (e.g., Hambleton & Swaminathan, 1985), testing the goodness-of-fit of IRT models is relevant to validating IRT applications. A common approach to assessing model-data fit is to compare the observed performance of individual items for various ability subgroups with the predicted performance under the chosen model. Several studies proposed utilizing a chi-square approximation of goodness-of-fit statistics, including Yen's (1981) $Q_1$ and Bock's (1972) $\chi^2$ indices. For polytomous IRT models, Muraki and Bock (1997) adopted a likelihood ratio chi-square statistic as a measure of fit for each item, and the sum of these chi-square statistics provides the likelihood ratio chi-square statistic for the whole test. However, a number of problem arise in using traditional chi-square statistics as tests for model-data fit in the IRT context. For example, theses chi-square type fit indices are hardly defensible against the criticism that they tend to be sensitive to test lengths and sample sizes leading to inflated Type I error rates.

Given the well-known problems associated with the traditional methods, alternative approaches to assessing IRT model-data fit have been proposed and successfully implemented in different applications (e.g., Donoghue & Hombo, 1999; Orlando & Thissen, 2000, 2003; Stone, 2000; Douglas & Cohen, 2001; Glas & Suárez Falcón, 2003; Sinharay, Johnson, & Stern, 2006). Donoghue and Hombo (1999) derived asymptotic distributions of the fit statistic under the assumption that item parameters are known. They demonstrated that the statistic behaved as expected under their restrictive assumption. Orlando and Thissen (2000) proposed fit statistics, $S-X^2$ and $S-G^2$ (discussed more fully later) based on joint likelihood distributions for each possible total score. They pointed out that the grouping of respondents in the traditional goodness-of-fit statistics is based on estimates that are model dependent, rather than on some observable statistics such as the number correct score. Alternatively, the $S-X^2$ and $S-G^2$ statistics are not dependent on the model-based trait estimates, but on directly observable frequencies and thus are solely a function of data. Stone (2000) used a resampling technique to determine the degrees of freedom and the scale factor for accounting for the uncertainty in estimated item parameters. For each replicated data set, the fit statistics, as well as the mean and the variance of the empirical distribution of the statistic are computed. Then the scale factor and the degrees of freedom are determined from the mean and the variance of the empirical distribution. On the other hand, Glas and Suárez Falcón (2003) proposed Lagrange Multiplier (LM) procedure to solve the uncertainty in the item parameter estimates and ability estimates in constructing the test statistic. More recently, Sinharay, Johnson, and Stern (2006) applied the Bayesian Posterior Predictive Model Checking (PPMC) procedure for evaluating model fit in IRT. Using the PPMC method with the Markov Chain Monte Carlo (MCMC) algorithm, they examined the performances of a number of discrepancy measures for assessing different aspects of fit of the common IRT models.

A number of simulation studies have been conducted to compare performances of these goodness-of-fit indices, and Type I error rates and power have been investigated. Stone and Zhang (2003) compared several different approaches to computing fit statistics including that of Bock (1972), Orlando and Thissen (2000), Stone (2000), and Donoghue and Hombo (1999). They found that the Orlando-Thissen and the Stone procedures had nominal Type I error rates, while unacceptably low or high Type I error rates were observed for the Donoghue-Hombo procedure and the Bock method. Glas and Suárez Falcón (2003) compared Orlando and Thissen's indices and their new proposal of the LM index with the usual log likelihood ratio $\chi^2$. They concluded that the Orlando-Thissen procedure had better overall characteristics than the LM index while the LM procedure was clearly superior to Yen's $Q_1$. A general finding from these comparative studies is that the new approaches (e.g., Orlando and Thissen's $S - X^2$ and $S - G^2$) appear to offer promising alternatives to traditional methods (e.g., Yen's $Q_1$ and Bock's $\chi^2$) for assessing goodness-of-fit of IRT models. Despite a number of applications of the alternative approaches to dichotomous response data, there is very little in the literature that addresses applications of goodness-of-fit indices such as $S-X^2$ to tests consisting of both dichotomous and polytomous items, which typically involve mixtures of different IRT models.

Therefore, the purpose of this study was to evaluate various IRT model combinations and calibration procedures for mixed-format tests, and to assess the appropriateness of the chosen IRT models and calibration procedure in terms of the model-data fit using traditional and alternative goodness-of-fit indices. For dichotomous items, one, two, and three parameter logistic (1PL, 2PL, and 3PL) models were considered. Among several polytomous models applicable to mixed format tests, Samejima's (1997) graded response (GR), Masters' (Masters & Wright, 1997) partial credit (PC), and Muraki's (1997) generalized partial credit (GPC) models were assumed for constructed-response items in this study. A variety of mixtures of these models were applied to real data sets and calibration was conducted both simultaneously and separately. The three fit statistics used in this study included PARSCALE's $G^2$ as a traditional index, and Orlando and Thissen's (2000) $S - X^2$ and $S - G^2$ as new fit indices. The original forms of $S - X^2$ and $S - G^2$ were extended for use with a mix of dichotomous and polytomous items in this study. Using the IRT models and fit indices under consideration, this study addressed the following questions:

- Which of the IRT model combinations leads to the best model-data fit for the given mixed format tests?

- How differently do the simultaneous and separate calibration procedures affect model-data fit for data containing multiple item types?

- Do the traditional and new fit indices provide comparable results in the evaluation of model-data fit?

# 2 Method

This section is organized as follows. First, data and dimensionality checking tools used in this study are described. Next, various mixtures of IRT models and estimation procedures for calibrating a mixture of two item types are presented. Finally, a brief mathematical presentation of the three item fit statistics is provided.

## 2.1 Data

The data sets used in this study consist of two parts, multiple-choice and constructed-response tests. For the multiple-choice tests, national standardization data for the Iowa Tests of Basic Skills (ITBS) Form M, Levels 13 and 14 (Hoover, Hieronymus, Frisbie, & Dunbar, 1996) for Reading Comprehension (Reading) and Math Problems and Diagrams (Math) were used. For the constructed-response tests, data from the Constructed-Response Supplement to the Iowa Tests (Hoover, Hieronymus, Frisbie, & Dunbar, 1996) were included in this study. These two tests were administered on different occasions; that is, students took a set of multiple-choice items first and then the same group of examinees took a set of constructed-response items as a supplement to the multiple-choice test. Although both item types were not completed in a single test administration, they were based on the same test specifications and designed to assess the same overall construct for each subject; i.e., reading comprehension and math problem solving ability. The Reading test for Level 13 (Grade 7) consists of 46 multiple-choice and 8 constructed-response items. Each multiple-choice item has five alternatives. Four constructed-response items are rated on a 3-point scale (0-2), and the other four constructed-response items are rated on a 2-point scale (0-1). Then Math test for Level 14 (Grade 8) consists of 36 multiple-choice and 13 constructed-response items. Each multiple-choice item in the Math test has five alternatives. Eleven constructed-response items are rated on a 3-point scale (0-2), and the other two constructed-response items are rated on a 2-point scale (0-1). Two separate data sets for Reading and Math contained item responses for 500 examinees.

## 2.2 Assessments of Unidimensionality

Because unidimensionality is one of the critical assumptions in the IRT models applied in this study, it is important to accurately assess the test structures prior to applying unidimensional IRT models. Furthermore, in deciding whether two item types can be calibrated together using unidimensional IRT models, the main issue to be considered is whether the two item types assess the same construct. In this study, two analytic procedures were used to assess the dimensional structure of each data set. First, the unidimensionality assumption was assessed by an examination of eigenvalues using factor analysis. When a rapid drop occurs between the first and second eigenvalues, unidimensionality can be approximately assumed. As an alternative approach, the Poly-DIMTEST (Li & Stout, 1995) procedure, in which exploratory factor analysis results are embedded to find item clusters (i.e., AT1, AT2, and PT), was used to test the hypothesis of unidimensionality. Stout (1990) argued that the traditional IRT assumptions of unidimensionality can be replaced by weaker assumptions of essential unidimensionality. Essential unidimensionality is defined as the presence of one dominant dimension and one or more minor dimensions that have relatively small influence on item scores (Stout, 1990). The Poly-DIMTEST program was applied to the mixed format data containing responses for both dichotomous and polytomous items.

## 2.3 IRT Model Combinations and Calibration Procedures

Examinees' responses to multiple-choice items were fitted by the 1PL, 2PL, and 3PL models. For constructed-response items, categorical responses were calibrated using the GR, PC, and GPC models. By combining these models, the following model mixtures (3 dichotomous × 3 polytomous models) were designed to analyze responses from the mixed format tests:

- 1PL and GR models,
- 1PL and PC models,
- 1PL and GPC models,
- 2PL and GR models,
- 2PL and PC models,
- 2PL and GPC models,
- 3PL and GR models,
- 3PL and PC models, and
- 3PL and GPC models.

For all the model combinations, both separate and simultaneous calibration procedures were considered. Thus, a total of 18 estimation conditions (9 model mixtures × 2 calibration procedures) were constructed and compared. The relative performances of the 18 conditions were investigated by three fit statistics at the individual item level as described in the next section.

The computer program PARSCALE (Muraki & Bock, 1997) was used to estimate item parameters for the mixed tests. Based on suggestions from previous studies (e.g., Chen, 1995; DeMars, 2005), the following PARSCALE options were employed to increase the precision of calibration. First, the number of EM cycles was increased to 200 and the number of Newton cycles was set to zero. Second, the number of quadrature points was set to 101. Third, prior distributions were used for item parameters. For estimation with GR or GPC models, $SSIGMA = 0.6$ was used as a prior for $a_j$ while $SSIGMA$ was set to a small value close to zero for the PC model; the $GPARM = 0.2$ option was included for the 3PL model.

## 2.4 Item Fit Statistics

To assess model-data fit for a mixture of multiple-choice and constructed-response items, PARSCALE's $G^2$ was chosen to represent the traditional chi-square type of fit statistics because this index is one of the most frequently used item fit statistics in practice. For alternative methods, the original forms of Orlando and Thissen's (2000) $S-X^2$ and $S-G^2$ for dichotomous items were extended to general forms for a mixture of the two item types.

### 2.4.1 PARSCALE's $G^2$

PPARSCALE provides a likelihood ratio chi-square statistic, $G^2$, as an item fit index. The $G^2$ statistic for item $j$ is

$$G_j^2 = 2\sum_{w=1}^{W_j}\sum_{k=0}^{K_j}\gamma_{wjk}\ln\frac{\gamma_{wjk}}{N_{wj}P_{jk}(\overline{\theta}_w)}, \tag{1}$$

where $k$ is the item score ranging from zero to the highest item score $K_j$; $W_j$ is the number of intervals left after neighboring intervals are merged, namely, the total number of groups of examinees. For each interval $w$, the interval mean, $\overline{\theta}_w$, the observed frequency, $\gamma_{wjk,}$ the total number of examinees, $N_{wj}$, and the fitted response function, $P_{jk}(\overline{\theta}_w)$, are computed. The degrees of freedom of $G^2$ is equal to the number of intervals, $W_j$, multiplied by $K_j-1$. Unlike some other fit indices such as Yen's (1981) $Q_1$ index, PARSCALE's $G^2$ does not adjust the degrees of freedom for uncertainty for either the item or proficiency parameters. As referenced in DeMars (2005), the rationale for the unadjusted degrees of freedom with the $G^2$ statistic can be found in Mislevy and Bock (1990). Mislevy and Bock (1990) asserted that the residuals are not under linear constraints so that there is no loss of degrees of freedom due to the fitting of the item parameters.

### 2.4.2 Orlando and Thissen's (2000) $S-X^2$ and $S-G^2$

In this study, extended forms of the $S-X^2$ and $S-G^2$ statistics were created for polytomous items by transforming the original forms of the indices for dichotomous items. The generalized forms of $S-X^2$ and $S-G^2$ for a mixture of the two item types are

$$S-X_j^2 = \sum_{h=h_{min}}^{h_{max}}\sum_{k_j=1}^{K_j}\frac{(O_{jkh}-E_{jkh})^2}{E_{jkh}}, \tag{2}$$

and

$$S-G_j^2 = \sum_{h=h_{min}}^{h_{max}}\sum_{k_j=1}^{K_j}O_{jkh}\ln\frac{O_{jkh}}{E_{jkh}}, \tag{3}$$

respectively. The observed proportions $(O_{jkh})$ and the expected proportions $(E_{jkh})$ for item $j$, category $k$, and number-correct score group $h$ are computed from the data. In these generalized formulas, category $K_j$ varies depending on the number of categories for each item. Using a recursive algorithm to obtain the joint likelihood of achieving summed score $h$, the expected proportions $(E_{jkh})$ are computed as

$$E_{jkh} = N_h \frac{\int P_{jk}(\theta) S^{*j}_{h-k} \phi(\theta) d\theta}{\int S_h \phi(\theta) d\theta}, \tag{4}$$

where $N_h$ is the number of examinees with score $h$; $P_{jk}$ is the item responses category function for item $j$ and category $k$; $S^{*j}_{h-k}$ is the posterior score distribution for score group $h - k$ for a scale without item $j$; $S_h$ is the posterior score distribution for score group $h$; and $\phi(\theta)$ is the population distribution of $\theta$. To compute $S_h$ and $S^{*j}_{h-k}$ for both dichotomous and polytomous items, the generalized recursive algorithm proposed by Thissen, Pommerich, Billeaud and Williams (1995) can be used. Because these statistics are based on the summed score distributions in which all items are answered, only complete cases in each data set were analyzed in calculating $S - X^2$ and $S - G^2$. Thus, 108 and 217 incomplete cases were excluded from the original data sets for Reading and Math, respectively. This occasionally led to reduced sample sizes in computing item fit.

# 3 Results

The results from the assessment of dimensionality as a preliminary analysis are presented first in this section. Then, the results of the principal analyses are summarized according to the three research questions regarding (1) the performance of the nine IRT model combinations, (2) simultaneous versus separate calibration procedures, and (3) the three indices assessing model-data fit. Graphical results of model fit are demonstrated to provide insights about misfitting items.

## 3.1 Assessments of Unidimensionality

Prior to applying unidimensional IRT models, the dimensional structures of the mixed format tests were examined to assess whether the two item types measure similar constructs. The results of the linear factor analyses and poly-DIMTEST procedures are presented in Table 1. The eigenvalues of the first factor were 9.61 and 8.35 for the Reading and Math tests, respectively, which were considerably greater than those for the remaining factors. These first components accounted for 17.79% and 17.04% of the total variances for Reading and Math, respectively. A rapid drop between the first and second eigenvalues suggested that there exists a single dominant dimension in each test. From this factor analysis approach, it seems reasonable to assume unidimensionality for each test. The results of poly-DIMTEST were consistent with the findings from factor analyses. That is, it can be seen that the unidimensionality hypothesis was retained from a $T$ statistic of -.79 and $p$-value of .787 for the Reading test. Similarly, the Math test was also assumed to be unidimensional from a $T$ statistic of .16 and a corresponding $p$-value of .436 as shown in Table 1. These results indicate that the dichotomous items were measuring the same construct as the polytomous items. Therefore, based on these two procedures, the unidimensionality assumption for each test seems to be highly defensible, making it reasonable to proceed with further analyses involving the use of unidimensional IRT models.

## 3.2 Performances of IRT Model Combinations

Item responses from the two mixed format tests for Reading and Math were calibrated under 18 conditions. For each condition, the three item fit statistics, $S-X^2$, $S-G^2$, and $G^2$, were computed at the individual item level. A significance level of .05 were used to examine misfitting items in this study. Significant fit statistics indicate that item parameters differ across the observed score groups and that the assumed model is not appropriate for the data.

Tables 2 and 3 present items flagged as misfitting at a significance level of .05 according to the nine model mixtures by the two calibration procedures for the Reading and Math tests, respectively. For the Reading test, items 1 through 46 are multiple-choice while the remaining items 47 through 54 are constructed-response. For the Math test, items 1 through 36 are multiple-choice while items 37 through 49 are constructed-response type. Figure 1 graphically summarizes numbers of misfitting items in the two tests. For the Reading test, estimation failed to converge in the conditions of the 1PL/PC and 1PL/GPC models with simultaneous calibration, and the three conditions involving the PC model with separate calibration. Item parameters for the 1PL/GR models with simultaneous calibration were estimated using 210 EM cycles to ensure convergence. Except for these six conditions, all estimations converged in PARSCALE within 35∼50 EM cycles using a set of priors for item parameters as described earlier. For the Math test summarized in Table 3, all estimations of the 18 conditions converged in PARSCALE within 35∼50 EM cycles using a set of prior for parameters as previously discussed.

In the Reading test, the overall pattern for the nine model mixtures in simultaneous calibration indicated that the 3PL/GPC and 2PL/GPC models led to the best fit to the given data set, while the 1PL/GR combination resulted in the worst fit consistently over the three indices. For separate calibration, the $G^2$ procedure yielded relatively more misfitting items. In contrast, the 2PL and 3PL combinations had no misfitting items according to $S-X^2$ and $S-G^2$ for separate calibration. Similarly, the math data set was best fit by the 3PL/GPC, 3PL/GR, 2PL/GPC, and 2PL/GR model combinations for both calibration methods. The proportion of misfitting items was larger when the PC model was combined with either the 2PL or 3PL model.

Comparing the three dichotomous models, model mixtures involving the 3PL model tended to exhibit fewer misfitting items across the three indices in the two tests regardless of the choice of the calibration procedure. Compared to the model mixtures involving the 2PL or 3PL model, a considerably larger number of items showed misfit in the conditions involving the 1PL model. For example, only one item was detected as misfitting by $G^2$ and $S-G^2$, and no item was detected as misfitting by $S-X^2$ with the 3PL/GPC model combination of simultaneous calibration for the Reading tests. By contrast, all items under the 1PL/GR model mixture turned out to be misfitting by $G^2$, and seven items out of 54 items were found as misfitting by $S-X^2$ and $S-G^2$ for the Reading data set. Comparing the three polytomous models, the GPC model exhibited fewer numbers of misfitting items than the other models consistently for all three fit indices. For example, no items showed misfit for any model combination involving the GPC model, whereas several items were found to be misfitting with the GR or PC models in simultaneous calibration using $S-X^2$ for Reading in Table 2.

## 3.3 Comparisons of Simultaneous and Separate Calibration Procedures

The fit of the items using various model combinations was compared under both simultaneous and separate calibrations. Compared to simultaneous calibration, separate calibration showed fewer misfitting items using $S-X^2$ and $S-G^2$ in the Reading test. That is, only two items were flagged as misfitting in the 1PL/GR and 1PL/GPC combinations, and all the other model conditions showed no misfits in separate calibration by the two new fit indices. In contrast, the calibration results for the Math test showed a slight tendency for the simultaneous calibration to have fewer misfitting items, particularly for the mixtures that included the 1PL model. Items flagged as misfitting matched between the two calibration conditions in the Reading data, but they did not match in the Math data. For example, in the Reading test summarized in Table 2, items 29 and 46 were misfitting for the 1PL/GR combination for both simultaneous and separate calibration procedures. By contrast, math item 11 was frequently detected as misfitting in simultaneous calibrations, but not in separate calibration. Comparing multiple-choice and constructed-response item sections in separate calibrations, higher percentages of misfit were observed in the calibrations of multiple-choice items over the two tests.

## 3.4 Comparisons of Item Fit Statistics

Three fit statistics were computed over the 18 estimation conditions for the two data sets. For both tests, the number of misfitting items using the $G^2$ statistic tended to be higher than the numbers of misfitting items found with the $S-X^2$ and $S-G^2$ indices, with a few exceptional cases. For example, all items exhibited misfit in the 1PL/GR combination using $G^2$ for the Reading data, while 7(13%) items showed misfit using $S-X^2$ and $S-G^2$. This different result for model-data fit using $G^2$ seems to be related to possibly unstable estimation with the particular model combination in PARSCALE. Because the statistic is based on the estimated proficiency distribution, unstable estimation of $\theta$ can result in an inflated degree of misfit. In fact, estimation with the 1PL/GR models required an increased number of EM cycles to obtain the converged solution. The fit of items with the 1PL/GR combination was clearly improved when they were calibrated with the more complex model mixtures, e.g., the 3PL/GR combination.

In terms of agreement among the three fit indices, the two new statistics showed considerable agreement for the misfitting items. Within each calibration method, items flagged for misfit by $S-G^2$ almost always overlapped with those detected by $S-X^2$ with only a few exceptions in the two tests. For example, reading items 49, 52, and 53 were observed as misfitting in the 3PL/PC model with the simultaneous calibration condition using $S-X^2$, and they were also identified as misfitting using $S-G^2$. By contrast, the $G^2$ index did not produce matching sets of misfit with the two new indices. As shown in Table 2, in the simultaneous calibration condition of the Reading test with the 3PL/PC combination, item 14 was flagged as misfitting using $G^2$, but not using $S-X^2$ and $S-G^2$. Conversely, items 49, 52, and 53 were found to be misfitting using $S-X^2$ and $S-G^2$, but not using $G^2$.

### 3.5 Graphical Examples of Item Fit Analysis

Graphical displays are useful in examining the discrepancy between observed and expected proportions correct (Swaminathan, Hambleton, & Rogers, 2007). Graphical examples of item fit analysis are shown in Figures 2 and 3. In these plots, differences between observed and expected frequencies for each summed score can be interpreted as degrees of model-data misfit. Figure 2 presents plots for dichotomously scored reading item 27 based on the three model combinations, 1PL/GR, 2PL/GR, and 3PL/GR. Among these model mixtures, the 2PL/GR and 3PL/GR combinations fit the data better than did the 1PL/GR combination since they showed narrower distances between observed and expected frequencies over the summed scores at each score point. Figure 2 showed the clear improvement in model fit of 2PL/GR over 1PL/GR, and the slight additional improvement of 3PL/GR over 2PL/GR. Similarly, Figure 3 enables comparing the fit of competing models for polytomously scored reading item 49 at each category score. These plots illustrated the gradual improvement in model fit of 3PL/GPC and 3PL/GR over 3PL/PC. This observation indicates that the model combination with the GPC or GR model, rather than the PC model, produced better model fit. Note that poor fit with 3PL/PC and good fit with 3PL/GPC and 3PL/GR for item 49 were reported in Table 2. In general, these plots suggest that the predictions of the test scores depend on IRT model combinations.

## 4 Summary and Discussion

Using real test examples, this study addressed three issues related to model-data fit in the mixed format tests. First, among various choices of the IRT model combinations for analyzing the mixed format tests, results of this study found that the 3PL or 2PL model with any of the polytomous models had fewer misfitting items than the 1PL model with the polytomous models. For the polytomous models, the GPC model tended to have slightly fewer misfitting items than the PC or GR models in some conditions. Second, a comparative look at the results from the two calibration procedures indicated that the choice of calibration procedure did not matter much in terms of model-data fit because the results were mixed. Finally, in comparing the three fit statistics, the traditional fit index, $G^2$, detected a somewhat different set of misfitting items for the two data sets, while there was considerable overlap using the two alternative statistics, $S-X^2$ and $S-G^2$ that were extended to more general forms for this study of mixtures of IRT models.

As for the performances of the various model mixtures, the results of this study are consistent with findings in Swaminathan, Hambleton, and Rogers (2007). A review of their study and the current study highlights the following common conclusions: For dichotomous items, the 3PL model fits the data better than the 2PL model, and both models fit the data better than the 1PL model. For polytomous items, the GPC model fit the data better overall than did the PC or GR model in some conditions. For the mixtures of dichotomous and polytomous items, the GPC model in combination with either the 3PL or 2PL model tended to fit the data best. Furthermore, a substantial improvement in overall model-data fit was found from the 1PL/PC or 1PL/GR to the 2PL/GPC model and the 3PL/GPC model combinations.

It should be noted that some model mixtures did not result in convergent solutions when estimating item parameters for the Reading test. Several factors could be associated with these estimation problems. Sample size may be a possible cause since IRT modeling requires a relatively large sample size, particularly with polytomous models. In this sense, the sample of 500 examinees might be considered to be too small for the purpose of evaluating model fit to the mixed format test data. On the other hand, failures in estimation can be viewed as an expected consequence of a bad fit according to the particular choice of model combinations. In fact, these problems occurred when the 1PL and PC models were fit to the Reading data set. Compared to the other models that allow item slope or discrimination parameters to vary, the 1PL and PC models have restrictive forms in that they include more constraints, i.e., a constant slope parameter, in their model specifications.

Misfit for some model combinations may be associated with innate properties of the given test data, e.g., dimensional structure of data sets and item characteristics. Since all of the IRT models considered in this study assumed unidimensionality, poor model-data fit in some conditions could be attributable to the fact that the assumption of unidimensionality was not fully satisfied even though preliminary analyses concluded that the assumption was met for both tests. Previous research indicates that dimensionality varies due to different conditions, such as content area and item format. Ercikan et al.(1998) also argued that the data structure of a mixed format test tends to be more multidimensional than a test involving single format items. This issue affects the appropriateness of the use of unidimensional measurement models as well as the interpretability of scores obtained by combining the two item types. Furthermore, Yen (1984) has shown that multidimensionality is expected to have a strong effect on the local independence of items as well as the fit of items to unidimensional IRT models. In this case, unidimensional models might not fit the data properly. Another explanation for poor fit could be related to the characteristics of the constructed-response items and the corresponding scoring rubrics. A detailed examination of misfit at the individual item level revealed that polytomous responses for some flagged constructed-response items became essentially dichotomous item responses because a certain category was not likely to occur as a response. Accordingly, these misfitting items had relatively low information and showed unusual patterns of item categorical response functions over the proficiency levels.

The results of the present study regarding the two calibration procedures suggested that the predictions of the test scores and model-data fit depend on the choice of a calibration method. Based on the two new fit statistics, simultaneous calibration led to slightly larger numbers of misfitting items than separate calibration for Reading, while converse results were observed for Math. Using real data sets, Ercikan et al. (1998) investigated psychometric properties (e.g., item fit, local item dependence, loss of information, and test information) for multiple-choice and constructed-response sections under the hypothesis that the two calibration procedures may perform differently if simultaneous calibration of the two item types leads to a loss of information. Ercikan et al. (1998) found that simultaneous calibration led to only negligibly small loss of information, and concluded that the results of their study support use of simultaneous calibration by combining scores from the two item types. Since both their study and the present study used only real data, useful future research would involve a comprehensive simulation study to evaluate the relative performance of the two calibration procedures in identifying misfitting items under various testing conditions.

The results for the goodness-of-fit statistics support the recent contention that the $S-X^2$ and $S-G^2$ statistics can be good alternatives for the evaluation of item fit, whereas the traditional $G^2$ statistic often draws implausible implications mainly with an unacceptably high proportion of items detected as misfitting, especially for a short test as discussed in Orlando and Thissen (2000, 2003), Glas and Suárez Falcón (2003), and Stone and Zhang (2003). However, the findings reported in this study should not be generalized beyond the specific measurement setting considered. Any decision on model-data fit has to account for the specific application of IRT modeling associated with various testing factors. In fact, administrative characteristics of the data sets used in this study could be more suitable to separate estimation since the two item types were given to the examinees on two different occasions. Furthermore, different proficiency distributions, sample sizes, estimation options, and subject matter could have affected the results. For example, with large sample sizes, even small deviations from model predictions can be statistically significant resulting in a large number of items flagged as misfitting. Therefore, the implications from this study can be informative as an empirical example of the topic of assessing IRT model-data fit for mixed format tests. To explore this issue further, it is recommended that future research include simulation studies. Particularly, the performances of the alternative item fit statistics need to be studied further for use with various test forms through simulation studies.

# 5 References

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, *28*, 77-92.

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

Chen, W. (1995). Estimation of item parameters for the three-parameter logistic model using the marginal likelihood of summed scores (Doctoral dissertation, University of North Carolina, 1995). *Dissertation Abstracts International*, *56/10-B*, 5825.

DeMars, C. E. (2005). Type I error rates for PARSCALE's fit index. *Educational and Psychological Measurement*, *65*, 42-50..

Donoghue, J. R., & Hombo, C. M. (1999). *Some asymptotic results on the distribution of an IRT measure of item fit.* Paper presented at the Annual Meeting of the Psychometric Society, Lawrence, KS.

Douglas, J. & Cohen, A. S. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, *25*, 234-243.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, *35*, 137-154.

Glas, C. A. W., & Suárez Falcón, J. C. (2003). A comparison of item-fit statistics for the three parameter logistic model. *Applied Psychological Measurement*, *27*, 87-106.

Hambleton, R. K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 3149). New York: Academic.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston MA: Kluwer-Nijhoff.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1996). *Iowa Tests of Basic Skills: Form M: Levels 13-14.* Itasca, IL: Riverside Publishing.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1996). *Constructed-Response Supplement to The Iowa Tests Form 1: Levels 13-14.* Itasca, IL: Riverside Publishing.

Kim, S. (2004). *Unidimensional IRT scale linking procedures for mixed-format tests and their robustness to multidimensionality.* Ph.D. Dissertation, The University of Iowa. Unpublished Manuscript.

Li, H.-H., & Stout, W. F. (1995, April). *Assessment of unidimensionality for mixed polytomous and dichotomous item data: Refinements of Poly-DIMTEST.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101-121). New York: Springer-Verlag.

Mislevy, R. J., & Bock, R. D. (1990). *PC-BILOG 3: Item analysis and test scoring with binary logistic models.* : Scientific Software, Inc.

Muraki, E., & Bock, R. D. (1997). *PARSCALE 3: IRT based item analysis and test scoring for rating-scale data.* Chicago, IL:Scientific Software International, Inc. [Computer Program].

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-185). New York: Springer-Verlag.

Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of $S - X^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*, 289-298.

Samejima, F. (1997). Graded response model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer-Verlag.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298-321.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, *37*, 58-75.

Stone, C. A., & Zhang, B. (2003). The effect of errors in estimating ability on goodness-of-fit tests for IRT models. *Journal of Educational Measurement*, *40*, 331-352.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293-325.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2007). Assessing the fit of item response theory models. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics*, *Vol. 26.* (pp. 683-718).

Sykes, R. C., & Yen, W. M. (2000). The scaling of mixed-item-format tests with the one-parameter and two-parameter partial credit. *Journal of Educational Measurement*, *37*, 221-244.

Thissen, D., Wainer, H., & Wang, X. B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement*, *31*, 113-123.

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

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett, & W. C. Ward (Eds), *Construction versus choice in cognitive measurement* (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed -response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*, 103-118.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245-262.

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*, 125-145.

Table 1: Results of Factor Analyses and Poly-DIMTEST Procedures

| | Factor Analysis | | | Poly-DIMTEST | |
|---|---|---|---|---|---|
| Test | Dimension | Eigenvalue | % of Variance | $T$ | $p$-value |
| Reading | 1 | 9.61 | 17.79 | -.79 | .787 |
| | 2 | 1.89 | 3.50 | | |
| | 3 | 1.56 | 2.89 | | |
| Math | 1 | 8.35 | 17.04 | .16 | .436 |
| | 2 | 1.83 | 3.73 | | |
| | 3 | 1.77 | 3.62 | | |

Table 2: Items Flagged as Misfitting in the Reading Test

| | $S-X^2$ | | $S-G^2$ | | $G^2$ | |
|---|---|---|---|---|---|---|
| Models | Misfit | N (%) | Misfit | N (%) | Misfit | N(%) |
| *Simultaneous Calibration* | | | | | | |
| 1PL/GR | 11, 23, 27, 29, 32, 39, 46 | 7(13.0) | 11, 23, 27, 29, 32, 39, 46 | 7(13.0) | All items | 54(100) |
| 1PL/PC | N/A | | N/A | | N/A | |
| 1PL/GPC | N/A | | N/A | | N/A | |
| 2PL/GR | 3, 39, 54 | 3(5.6) | 3, 54 | 2(3.7) | 5-7, 14, 15, 18, 26, 28, 35, 36, 38, 44, 54 | 13(24.1) |
| 2PL/PC | 49, 52, 53 | 3(5.6) | 3, 49, 52, 53 | 4(7.4) | 14, 38, 44 | 3(5.6) |
| 2PL/GPC | | 0(0.0) | 3 | 1(1.9) | 14, 38, 44 | 3(5.6) |
| 3PL/GR | 3, 39, 54 | 3(5.6) | 3, 54 | 2(3.7) | 14, 15, 38, 52, 54 | 5(9.3) |
| 3PL/PC | 49, 52, 53 | 3(5.6) | 3, 49, 52, 53 | 4(7.4) | 14 | 1(1.9) |
| 3PL/GPC | | 0(0.0) | 3 | 1(1.9) | 14 | 1(1.9) |
| *Separate Calibration* | | | | | | |
| 1PL/GR | 29, 46 | 2(3.7) | 6, 29, 46 | 3(5.6) | 6, 26, 28, 29, 35, 46, 49-51, 53 | 10(18.5) |
| 1PL/PC | N/A | | N/A | | N/A | |
| 1PL/GPC | 29, 46 | 2(3.7) | 6, 29, 46 | 3(5.6) | 6, 26, 28, 29, 35, 46, 49-51, 53 | 10(18.5) |
| 2PL/GR | | 0(0.0) | | 0(0.0) | 38, 44, 49-51, 53 | 6(11.1) |
| 2PL/PC | N/A | | N/A | | N/A | |
| 2PL/GPC | | 0(0.0) | | 0(0.0) | 38, 44 49-51, 53 | 6(11.1) |
| 3PL/GR | | 0(0.0) | | 0(0.0) | 5, 14, 38, 49-51, 53 | 7(13.0) |
| 3PL/PC | N/A | | N/A | | N/A | |
| 3PL/GPC | | 0(0.0) | | 0(0.0) | 5, 14, 38, 49-51, 53 | 7(13.0) |

*Note.* N/A indicates that parameter estimates were not available.

Table 3: Items Flagged as Misfitting in the Math Test

| | $S-X^2$ | | $S-G^2$ | | $G^2$ | |
|---|---|---|---|---|---|---|
| Models | Misfit | N(%) | Misfit | N(%) | Misfit | N(%) |
| *Simultaneous Calibration* | | | | | | |
| 1PL/GR | 11 | 1(2.0) | 5, 11, 12 | 3(6.1) | 2, 4, 5, 9, 10, 14, 18, 30 | 8(16.3) |
| 1PL/PC | 11, 12, 39 | 3(6.1) | 5, 11, 12, 39 | 4(8.2) | 4, 5, 9, 10, 14, 18, 19, 30 | 8(16.3) |
| 1PL/GPC | 11 | 1(2.0) | 5, 11, 12 | 3(6.1) | 4, 5, 9, 10, 14, 18, 19, 30 | 8(16.3) |
| 2PL/GR | | 0(0.0) | 11, 12 | 2(4.1) | 12, 19, 23 | 3(6.1) |
| 2PL/PC | 39 | 1(2.0) | 11, 12 | 2(4.1) | 19 | 1(2.0) |
| 2PL/GPC | | 0(0.0) | 11, 12 | 2(4.1) | 19 | 1(2.0) |
| 3PL/GR | | 0(0.0) | 11, 12 | 2(4.1) | | 0(0.0) |
| 3PL/PC | 39 | 1(2.0) | | 0(0.0) | | 0(0.0) |
| 3PL/GPC | | 0(0.0) | 11, 12 | 2(4.1) | | 0(0.0) |
| *Separate Calibration* | | | | | | |
| 1PL/GR | 4, 14, 18, 19, 30, 33 | 6(12.2) | 4, 10, 14, 18, 19, 30, 33 | 7(14.3) | 1, 4-6, 9, 10, 14, 15, 18, 23, 25, 26, 30, 32, 35 | 15(30.6) |
| 1PL/PC | 4, 14, 18, 19, 30, 33, 42 | 7(14.3) | 4, 10, 14, 18, 19, 30, 33 | 7(14.3) | 1, 4-6, 9, 10, 14, 15, 18, 23, 25, 26, 30, 32, 35 | 15(30.6) |
| 1PL/GPC | 4, 14, 18 19, 30, 33 | 6(12.2) | 4, 10, 14, 18, 19, 30, 33 | 7(14.3) | 1, 4-6, 9, 10, 14, 15, 18, 23, 25, 26, 30, 32, 35 | 15(30.6) |
| 2PL/GR | 33 | 1(2.0) | 19, 33 | 2(4.1) | 12, 14, 19 | 3(6.1) |
| 2PL/PC | 33, 42 | 2(4.1) | 19, 33 | 2(4.1) | 12, 14, 19 | 3(6.1) |
| 2PL/GPC | 33 | 1(2.0) | 19, 33 | 2(4.1) | 12, 14, 19 | 3(6.1) |
| 3PL/GR | 33 | 1(2.0) | 33 | 1(2.0) | 1, 6, 10 | 3(6.1) |
| 3PL/PC | 33, 42 | 2(4.1) | 33 | 1(2.0) | 1, 6, 10 | 3(6.1) |
| 3PL/GPC | 33 | 1(2.0) | 33 | 1(2.0) | 1, 6, 10 | 3(6.1) |

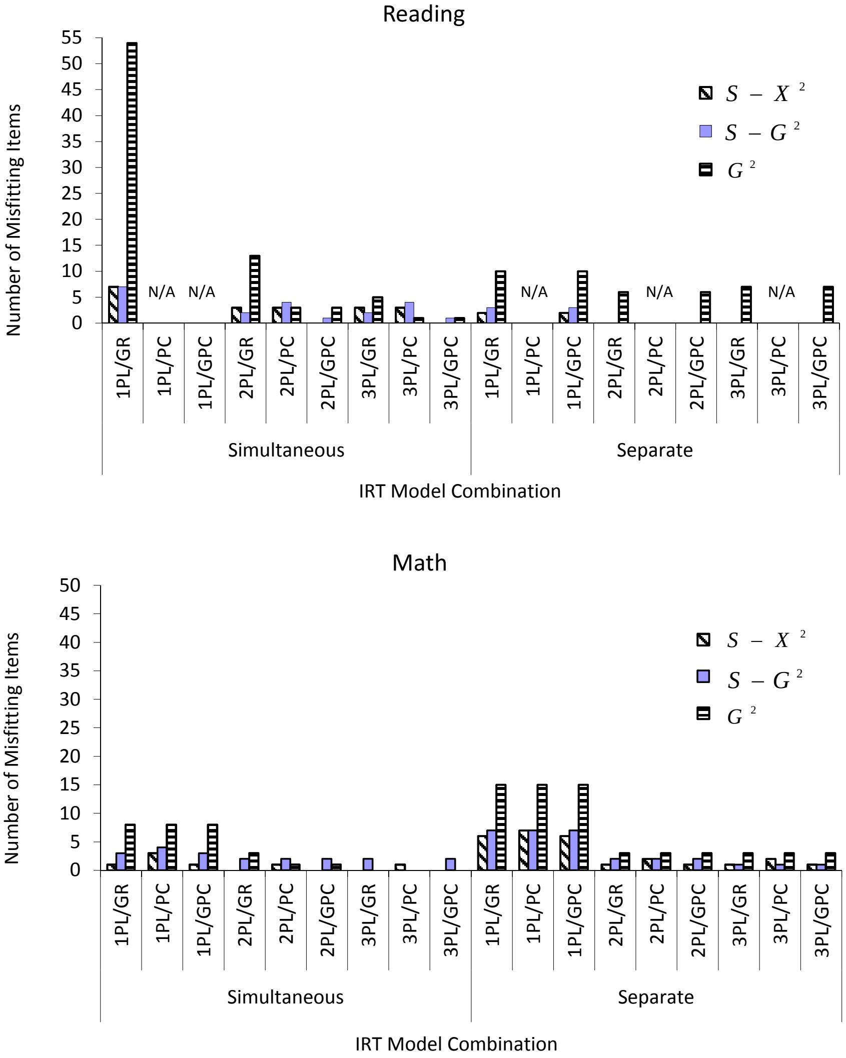

Figure 1: Number of Misfitting Items for Reading and Math

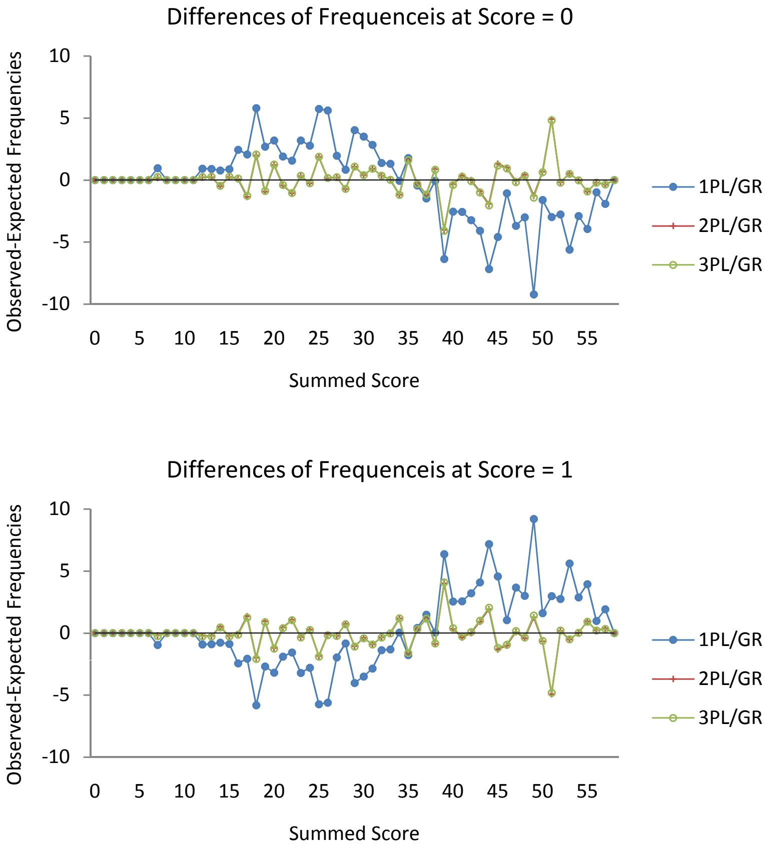

Figure 2: Differences of Observed and Expected Frequencies for Each Summed Score of Reading Item 27 (Dichotomously Scored Item)

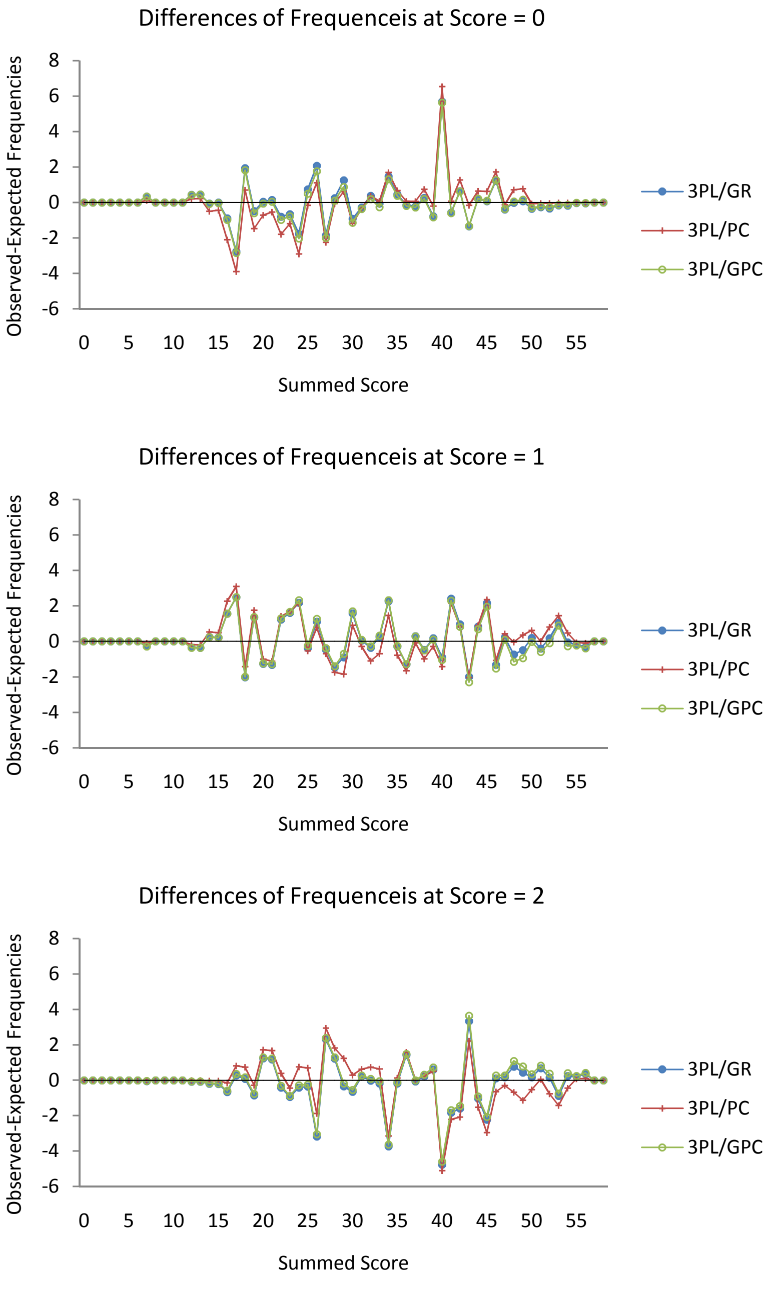

Figure 3: Differences of Observed and Expected Frequencies for Each Summed Score of Reading Item 49 (Polytomously Scored Item)