The Effect of Nonignorable Missing Data in Computerized Adaptive Test on Item Fit Statistics for Polytomous Item Response Models

Shudong Wang
Northwest Evaluation Association

Hong Jiao
University of Maryland

Yun Xiang
Northwest Evaluation Association

Paper presented at the annual meeting of the National Council on Measurement in Education. April 27-30, 2013, San Francisco, CA.

Send correspondence to:

Shudong Wang
Northwest Evaluation Association (NWEA)
121 NW Everett St.
Portland, OR 97209
Shudong.Wang@NWEA.org

Introduction

For both linear and adaptive tests, it is crucial to evaluate mode-data fit because the goodness-of-fit (GOF) of item response theory (IRT) models are relevant to any purpose of a test. To date, all item fit statistics are derived based on linear tests and almost all studies have been done in the context of linear testing. These studies are conducted based on assumptions under regular conditions for fixed test forms, such as no missing responses and normal distribution of unidimensional ability for a population.

Because sample and item invariance properties of item response theory (IRT) heavily rely on how well model and data fit in current testing practices, evaluating the GOF of a model by examining item and person fit statistics becomes an important part of operation procedure in validating the appropriateness of IRT model (Wells & Bolt, 2004; Chon, Lee, & Dunbar, 2010; Hambleton & Han, 2004; Sinharay & Lu, 2008; Stone & Zhang, 2003). A general approach to evaluating GOF involves the comparison between observed and model-predicted distributions for various ability subgroups using chi-square fit statistics. Among many procedures to assess GOF of dichotomous and polytomous IRT models at item level (Bock, 1972; Douglas & Cohen, 2001; Glas & Suarez-Falcon, 2003; McKinley & Mills, 1985; Orlando & Thissen, 2000, 2003; Sinharay, 2006; Stone, 2000; Yen, 1981), the GOF statistics can be classified as traditional classical fit statistics, such as PARSCALE's or Bock's (Murak & Bock, 2003) G^2, and Yen's (Yen 1981) Q_1 indexes; and Stone's (Stone, 2000) pseudo-observed score fit statistics χ^{2*} and G^{2*}, in which both type fit statistics use model-based theta estimates to obtain the observed proportions. The alternative item fit statistics, such as $S-X^2$ and $S-G^2$ statistics (Orlando & Thissen, 2000, 2003), are based on joint likelihood distributions for each possible summed raw score. The performance of both types of item fit statistics applied to mixed dichotomous and polytomous items have been evaluated by many researchers (Chon, Lee, & Dunbar, 2010; Dodeen, 2004; Sinharay & Lu, 2008). However, all these studies are conducted under normal conditions of linear (fixed form) test assumptions, such as no missing responses and normal distribution of ability population with unidimensionality.

Compared to the sample data of item calibration for a linear test, the conditions to collect item calibration sample data for a computerized adaptive test (CAT) are usually less ideal because of the intrinsic nature of the CAT test—i.e., the restriction of the ability range and sparseness of data matrices (Ban, Hanson, Yi & Harris, 2002; DeMars, 2002; Glas, 2000; Glas &

Pimentel, 2008; Harmes, Parshall & Kromrey, 2003)—and the problems of restriction range and sparse matrix in CAT data is essentially a problem of missing data. According to Rubin's (1976) missing data mechanisms, educational data can be classified as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). Within the latent variable modeling framework (Muthén, Asparouhov, Hunter, & Leuchter, 2011), if data missing is related to observed variables, then it can be MAR; if data missing is related to latent variables, such as student achievement ability, then it is MNAR, and such missing data refers to non-ignorable missing data. The type of missing data in CAT is MNAR. Tables 1 and 2 show examples of missing data in linear and CAT tests. Because $S-X^2$ and $S-G^2$ statistics are based on sum scores and this type of statistic cannot be used for CAT data where each student gets different items, this study only focuses on the performance of χ^{2*} and G^{2*} statistics.

The impact of CAT data on item parameter estimation has been studied by many researchers. Wainer and Mislevy (2000) and van der Linden and Glas (2000) investigated capitalization of item calibration errors in CAT; other studies (Lord & Wingersky, 1984; Thissen & Wainer, 1982) show that precision of item parameter estimation correlates directly with the distribution ability of examinees used for calibration. The performances of fit statistics for linear tests have been extensively examined under different conditions such as sample size, sample distribution, test length, IRT model type, and mixed item format. A few researches have examined the impact of MNAR absence in CAT data on the performance of these GOF statistics. CAT has been used in licensure and certification for decades and nowadays, CAT is becoming more popular in medical and educational tests. Right now, Oregon, Delaware, and Idaho use CAT in their state assessments, and several other states (Georgia, Hawaii, Maryland, North Carolina, South Dakota, Utah, and Virginia) and the Smarter Balanced Assessment Consortium are in various stages of CAT development. Other than dichotomously scored items, all state assessment programs require polytomous scored items in their tests and currently, most states use one or two of following dichotomous and polytomous IRT models in their state programs: (1) Rasch model (Rasch, 1960), (2) Three-parameter logistic model (3PL, Lord & Novick, 1968), (3) Samejima's (1969) graded response model (GRM), (4) Muraki's (1992) generalized partial credit model (GPCM), (5) Master's (1982) partial credit model (PCM), and (6) Andrich's (1978) rating scale model (RSM). In general, the advantages of a polytomous model are (a) the amount of item information provided by a polytomously scored item is greater than that from a dichotomously

scored item (Baker, 1992; Bock, 1972; Sympson, 1983; Thissen & Steinberg, 1984, Samejima, 1969); (b) the rate of detecting mismeasured examinees using a polytomously scored item is greater than it is when using a dichotomously scored item.

Given the increasing popularity of CAT in statewide assessments along with the wide adoption of polytomous IRT models for performance-based items, particularly in the implementation of the Common Core State Standards (CCSS), there is a pressing need to evaluate the performance of some commonly-used fit statistics for polytomously scored items, which were developed based on linear tests but have been employed to the adaptive testing. Through a simulation, this study examines the impact of missing data on the item fit statistics, χ^{2*} and G^{2*}, between a linear test and a computerized adaptive test based on IRT.

Method

IRT data-model fit using fit statistics

The basic idea of testing IRT data-model fit using fit statistics is to compare expected and observed frequencies of item category (either dichotomous or polychromous) responses for different IRT ability (theta) scores. Traditional IRT fit statistics treat estimated theta as observed scores, and major steps to calculate traditional IRT fit statistics involve

1) Grouping ability to approximate continuous theta distribution (for example, 10 groups from -4 to 4). How the intervals are created and how many intervals are created are arbitrary.
2) For any ability groups, getting observed score distribution and expected score distribution using IRT model, item parameter estimate, and midpoint of theta level of subgroup.
3) Comparing observed and expected distributions and examining the residual for each item. For example,
   i) Chi-square (or likely ratio of chi-square) based on the GOF statistics:

$$\chi^2 = \sum_K \sum_J n_k \frac{(O_{kj} - E_{kj})^2}{E_{kj}}, \quad (1)$$

   Where *k* and *j* represent the ability subgroup of theta and score response, respectively, O_{kj} and E_{kj} are the observed and expected proportions for ability subgroup *k* and score response *j*, respectively, and n_k is the frequency of individuals in subgroup *k*.

Yen's Q_1 (1981) is for dichotomously scored item χ^2, and Bock's χ_B^2 (1972) is similar, except *k* may vary.

ii) The standardized residual based on the GOF statistics (Hambleton & Han, 2004):

$$z_{kj} = \frac{(O_{kj} - E_{kj})}{\frac{\sqrt{E_{kj}(1-E_{kj})}}{n_k}}, \quad (2)$$

The fundamental problem using theta values to group frequencies is that theta values are never directly observed, because theta is latent-variable. Two different ways to deal with this problem are (1) using raw score and (2) using pseudo-observed score distribution.

(1) Using total raw score instead of theta in fit statistics. For example, for *S*-χ^2 and *S*-G^2 fit statistics proposed by Orlando and Thissen (2000, 2003) for dichotomous items, the examinees are divided into *n* groups based on total raw score. The χ^2 and G^2 for polytomous items have the form for item *i*:

$$S - \chi_i^2 = \sum_{k=k_{min}}^{k_{max}} \sum_{c=1}^{C_i} n_h \frac{(O_{ick} - E_{ick})^2}{E_{ick}}, \quad (3)$$

and

$$S - G_i^2 = 2 \sum_{k=k_{min}}^{k_{max}} \sum_{c=1}^{C_i} O_{ick} \ln(\frac{O_{ick}}{E_{ick}}), \quad (4)$$

where n_k is the number of examinees in raw score group *k*; O_{ick} and E_{ick} are observed and expected proportions, respectively, for item *i*, category *c* and summed score group *k*. The join likelihood of achieving a summed raw score *k* can be obtained by using recursive algorithm, and the expected proportions are computed by

$$E_{ick} = \frac{\int P_{ic}(\theta) S_{k-c}^{*i} \emptyset(\theta) d\theta}{\int S_k \emptyset(\theta) d\theta}, \quad (5)$$

where $P_{ic}(\theta)$ is the item response category function for category *c* of item *i*; S_{k-c}^{*i} is the posterior score distribution for score group *k*-*c* for a scale without item *i*; $S_k$is the posterior score distribution for score group *k* and $\emptyset(\theta)$ is the population distribution of ability θ.

(2) χ^{2*} and G^{2*} fit statistics use pseudo-observed score distribution (PSSD) for a limited number of discrete ability θ points to take precision of theta estimate into account (Stone, 2000; Stone, Mislevy, & Mazzeo, 1994). The PSSD is the posterior expectations that classify each examinee into several cells of the item fit table based on their entire distribution of posterior expectations of ability θ. The χ^{2*} and G^{2*} for item *i* are computed as

$$\chi_i^{2*} = \sum_{k=1}^{K_i} \sum_{c=1}^{C_i} \frac{r_{i.k}[(r_{ick}/r_{i.k}) - E_{ick}]^2}{E_{ick}}, \quad (6)$$

and

$$G_i^{2*} = 2 \sum_{k=1}^{K_i} \sum_{c=1}^{C_i} \frac{r_{ick}}{r_{i.k}} \ln[\frac{r_{ick}/r_{i.k}}{E_{ick}}], \quad (7)$$

where r_{ick} and E_{ick} are the pseudo-count and expected proportion, respectively, for categorical response c of item *i* at ability level *k* ($\theta = k$), and $r_{i.k}$ represents the posterior expectation of the number of attempts at ability level *k* for item *i*. The difference of theta-based fit statistics between traditional and the PSSD is that an examinee's contribution to the item fit is distribution over θ levels rather than restricting the contribution to a single cell based on a point estimate of θ. According to Stone (2000), the more imprecisely θ is estimated, the wider the posterior expectations are distributed across the θ range. The null hypothesis of fit statistics H_0 for a given item is:

$$H_0: \pi_{kc} = P_c(\theta_k),$$

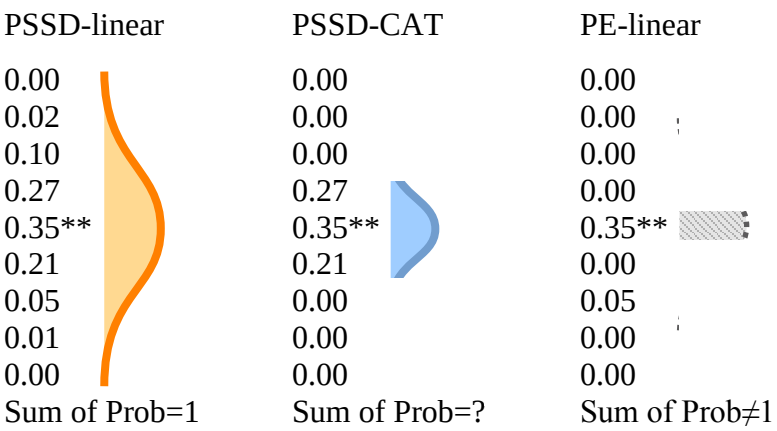
where π_{kc} is proportion of individuals from the population scoring at score level *c* and ability level *k*, and $P_c(\theta_k)$ is probability of individuals from the population scoring at score level *c* and ability level *k* from a given model.

Stone (2000) provided an example of distribution of pseudo counts for three students responding with scores of 0, 3 and 4 to a given item in linear test:

Stone's example

θ group	0	1	2	3	4
-1.90	0.00				
-1.47	0.02				
-1.05	0.10				
-0.63	0.27			0.00	
-0.21	0.35**			0.03	
+0.21	0.21			0.19	
+0.63	0.05			0.41**	0.00
+1.05	0.01			0.29	0.05
+1.47	0.00			0.07	0.23
+1.90				0.00	0.37**
+2.32					0.25
+2.74					0.08
+3.16					0.01

The difference of posterior expectations of theta for a point estimate (PE) from a linear test (PE-linear), PSSD from a linear test (PSSD-linear; sum of the probabilities for given examinee is 1), and posterior expectations of theta from a CAT test (PSSD-CAT; sum of the probabilities for given examinee may or may not be 1) can be illustrated for score 0 of a given item:



The impact of the difference between the PSSD-linear and the PSSD-CAT on χ^{2*} and G^{2*} is our major focus for this study.

Design of Simulation Study

The primary goal of this design is to maximize generalizability and replicability of research results. The polytomous IRT models investigated are the generalized partial credit model (GPCM, Muraki, 1992) and the graded response model (GRM, Samejima).

Table 3 lists all four variables manipulated in the study and their IRT models, test length (20, 40, 100), testing algorithm (linear and CAT), and missing rate (MR 0, 0.80 and 0.60). The reason to choose the highest MR around 0.8 is to set a low boundary of MR in the study. In practice, most real CAT programs have MR larger than 0.8. For example, in the NWEA MAP reading CAT test, the usual MR reaches above 0.90 (Wang & Harris, 2011). The MR is not an independent variable because MR is function of test item length and bank size:

$$MR=1-\frac{\text{Test Length}}{\text{Item Bank Size}}. \qquad (1)$$

A sample size of 10,000 is used across all conditions. The dependent variables are empirical Type I error rates for two fit statistics χ^{2*} and G^{2*}. The performance of fit statistics based on a linear test will be used as the baseline so the performance of fit statistics based on CAT can be compared to these baseline results. A total of 50 replications for each design condition will be conducted in this study. Because pseudo-observations are not independent, which means probability for each examinee contributes to multiple groups, and one of the family of chi-squared distributions cannot be assumed (Stone, 2000), the fit statistic is a scaled chi-squared random variable (Stone, Ankenmann, Lane & Liu, 1993). Resampling procedures can be used to obtain estimates of the scaling factors and effective *df* (γ and ν). According to Stone (2000), the Monte Carlo re-sampling procedure for estimating scaling corrections is used to approximate a null chi-square distribution. The resampling procedure involves the following:

1. Given item parameter estimates from scaling the observed test data, calculate the fit statistic using the posterior expectations;
2. Generate K random samples under H_o using the item parameter estimates from the original data and an assumed ability distribution (N(0,1));
3. Compute the item fit statistics using the posterior expectations for each k^{th} simulated sample;
4. From the empirical sampling distribution under H_o (K fit statistics), compute estimates of γ and ν; and
5. Rescale the fit statistic using γ and compare the rescaled fit statistic to a chi-squared distribution with ν (*df*).

In this study, the number of replication of resampling is 100 across all conditions. A total of 30 replications for each design condition will be conducted in this study.

Generation of Item and Person Parameters

Table 4 presents information about distributions of person-ability parameters, item discrimination (slope) and categorical (step) parameters.

Data Analysis

Both linear and CAT response data were generated using SAS (SAS Institute Inc., 2008).

All CAT tests are fixed test-length tests. The item-selection procedure is a maximum information-selection method among a given group of items that have item locations parameter values ranging from -0.2 to 0.2 logit around any provisional ability estimate during the CAT test. The ability estimation method used in this study is the expected *a posteriori* (EAP: Bock & Aitkin, 1981) method. For the linear test, test length is used as size item bank, and CAT simulation runs through the whole bank to get linear test responses. For example, for a 20 item test, the bank size is set to 20 and run a CAT test length of 20 items. After generating both linear and CAT responses, the responses based on GPCM and GRM models are calibrated using PARSCALE. The calibrated item parameters and generated responses for both models are used as input for computing fit statistics.

The fit statistics χ^{2*} and G^{2*} and resampling are computed using the SAS macro IRTFIT (Bjorner, Smith, Stone, & Sun, 2007). Because Stone's fit statistics is the theta-based method, the likelihood for theta that is necessary to calculate the pseudo-counts can be computed even if some items are unanswered, because IRTFIT will use information for all items that are not-missing. If the observed count of a given item for some ability level is very low, IRTFIT will not output fit statistics for that item, and we label that kind of item as not-used items in the item bank.

Results

Table 5 shows the Type I error rates (α) of χ^{2*} and G^{2*} across different simulation conditions. α is the number of flagged misfit items under a given condition. Since the number of replication of resampling is set to 100, the number of flagged misfits is also the percentage of flagged misfit items under given condition. As shown in Table 5, in general, for a fixed-length test that has a 100% ratio of test length over bank size, χ^{2*} and G^{2*} can be obtained for all items and this is reflected in the column labeled " % Average Number Item Used"; for CAT test, χ^{2*} and G^{2*} can be obtained only for part of the items in item bank. For example, based on the GPCM and for test length 20 and bank size 100, fit statistics χ^{2*} can be obtained for only 53% of items, and the rest have a very low count in cells to calculate fit statistics. Although only partial items in the item bank can be evaluated using fit statistics, the number of used items (53 items) is still greater than that of the test length (20 items), and "% Average Number Item Used" is also greater that the ratio of test length over bank size.

As can be seen in Table 5, the average Type I error rate increases as test length increases for a linear test, and this is expected because as the number of items increases, the chance of getting misfit items increases. The interesting finding for this study is that the average Type I error rate decreases for a CAT test compared to a linear test. For example, for an item bank of 100, both Type I error rates for test lengths 20 and 40 are smaller than those of a 100 test-length test, which is a linear test. However, this reduction in Type I error rates could be due to a shorter test compared to 100 items. The results of Type I error rates across models show that items based on a GPCM model have better fits than those of GRM, and this could come from the fact that GRM does not allow step reverse. In this study, the simulated item parameters for both GPCM and GRM come from the same distributions (see Table 4), and all the distances between step parameters are one logit. Although the chances that easy step parameters have larger logit values than any adjacent hard-step parameters are very slim, it still exists. This is not a problem for GPCM, but it could cause problems for GRM.

Educational Importance of the Study

Due to the advantages of CAT over linear tests in state and large scale assessments, CAT is gaining popularity in statewide assessment. Model-data fit is a very critical step in CAT applications like CAT item development and scoring. One of the direct consequences of failing to detect GOF of items in CAT is being unable to estimate a student's ability correctly. The errors in student achievement ability estimates could result in misclassifying students in educational learning, instruction and evaluation decisions. The majority of previous studies on the GOF of IRT models focused on linear tests, and all GOF statistics used so far were developed for linear tests, so little is known about whether GOF statistics results based on linear tests can be generalized to CAT. The results from this research provide empirical evidence regarding the effects of CAT data on GOF statistics and can be used to guide practitioners about performance of GOF statistics in CAT environments. First, we recommend not applying the same set of rules for item fit to CAT and linear tests. In either standard-alone or embedded-field CAT tests, we should not expect all field-test items to be used to fit IRT models. Hence the attrition rate (percentage of field-testing items that will not be used as operational items) due to CAT algorithm alone (having nothing to do with item quality) will be higher than that in linear field testing. Second, because of the intrinsic nature of CAT (i.e., the restriction of ability range of

examinees), the responses collected from CAT have impact item fit results compared with linear tests. Hence only selected item fit statistics that use theta as a grouping variable can be used to evaluate item model fit. The results from this study show that CAT field testing design needs more prudent thinking than that of linear tests.

Table 1: Example of Missing Data (Dichotomous Item Responses) from a Linear Test with Test Length = 5 and Number of Persons = 20*

	Item						
	Sub-content $_1$		Sub-content 2			Sub-Total$_1$	Sob-Total$_2$
Person	I_1	I_2	II_1	II_2	II_3	RS1	RS2
P1	1	1	1	1	0	2	2
P2	1	.	1	0	.	1	1
P3	1	0	1	1	0	1	2
P4	1	1	0	1	0	2	1
P5	.	1	1	0	1	1	2
P6	1	1	0	0	0	2	0
P7	1	0	1	1	0	1	2
P8	1	1	0	1	0	2	1
P9	1	.	1	.	1	1	2
P10	1	1	0	0	0	2	0
P11	1	1	1	0	.	2	1
P12	1	0	1	1	0	1	2
P13	1	1	0	1	0	2	1
P14	0	1	.	0	1	1	2
P15	1	1	0	0	0	2	0
P16	.	0	1	.	0	1	1
P17	1	1	0	1	0	2	1
P18	1	1	1	1	0	2	2
P19	1	.	1	0	0	1	1
P20	1	0	1	1	0	1	2

*: "." represents missing

Table 2: Example of Missing Data (Dichotomous Item Responses) Sorted by Person Ability (from Low to High) and Item Difficulty (from Easy to Hard) from a CAT Test (Due to Test Design) with Test Length = 5 out of Item Bank Size= 30 and Number of Persons = 20*

	Item																													Sub-Total$_1$	Sob-Total$_2$
	Sub-Content $_1$ + Sub-Content $_2$																														
	I	I	I	I	I	I	I	I	I	II	I	I	I	I	II	II	II	II	I	II	II	II	II	II	II	II	II	II	II		
Person	1	2	3	4	14	6	7	8	12	13	11	9	10	5	15	16	17	18	3	26	27	28	29	24	25	19	20	21	22	30	RS1	RS2
P15	1	.	0	1	0	0	.	.	.	.	.	.	.	.	.							.	.	.	.	.	.	.	.	.	2	0
P6	.	1	1	0	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	2
P10	.	.	1	.	1	0	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	2
P4	.	.	.	0	1	1	1	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	2
P5	.	.	.	.	.	1	0	1	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	0
P6	.	.	.	.	.	.	.	1	.	0	0	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	0
P7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	1
P8	.	.	.	.	.	.	.	.	.	.	.	1	0	1	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	1
P9	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	0	0	.	.	.	.	.	.	.	.	.	.	.	.	.	2	1
P10	.	.	.	.	.	.	.	.	.	.	.	.	.	1	0	0	1	1	.	.	.	.	.	.	.	.	.	.	.	.	1	2
P11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	0	1	0	.	.	.	.	.	.	.	.	.	.	1	2
P12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	0	1	0	.	.	.	.	.	.	.	.	3	0
P13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1		1	0	1	0	.	.	.	.	.	.	.	.	2	1
P14	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	1	1	1		0	.	.	.	.	.	.	.	1	2
P15		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	0	.	1		0	.	.	.	.	.	.	1	2
P16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	0	1	0	.	.	.	.	.	.	1	2
P17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	0	1	0	.	.	.	.	.	3	0
P18	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	0	1	0	.	.	.	.	3	0
P19	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	1	1	0		0	.	.	.	2	0
P20	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	0	1		0	3	0

*: "." represents missing

Table 3: Research Design (Sample Size = 10,000)

Design	IRT Model	Test Length	Testing Algorithm	Item Bank Size	Missing Rate (%) (1-Test Length/Bank Size)
1	GPCM	20	Linear	20*	0
	GRM	20	Linear	20*	0
2	GPCM	40	Linear	40*	0
	GRM	40	Linear	40*	0
3	GPCM	20	CAT	100	80
	GRM	20	CAT	100	80
4	GPCM	40	CAT	100	60
	GRM	40	CAT	100	60
5	GPCM	100	CAT	100	0
	GRM	100	CAT	100	0

*: Items for linear tests are drawn from generated CAT item banks, and then matched to items used for CAT tests so that performances of these items can be compared between linear and CAT tests.

Table 4: Generated Parameter Distributions of Models (Sample Size = 10,000)

Model	θ	a	b	d_1	d_2	d_3	d_4
GPCM	$N(0,1)$	Log($N(0,0.4)$)		$N(-1.5,0.5)$	$N(-0.5, 0.5)$	$N(0.5, 0.5)$	$N(1.5, 0.5)$
GRM	$N(0,1)$	Log($N(0,0.4)$)		$N(-1.5,0.5)$	$N(-0.5, 0.5)$	$N(0.5, 0.5)$	$N(1.5, 0.5)$

Table 5: Average Type I Error Rates (α) of Goodness-of-Fit Statistics X^{2*} and G^{2*} over 30 Replications

Model	Fit Statistics	Bank Size	Test Length	Average Number of Items Used	% Average Number of Items Used	Missing Rate (%)	Ratio of Test Length/ Bank Size	Average Type I Error Rates (α)
GPCM	X^{2*}	20	20	20.00	100.00	00.00	1.00	2.00
		40	40	40.00	100.00	00.00	1.00	4.30
		100	20	52.75	0.53	80.00	0.20	1.25
		100	40	76.89	0.77	60.00	0.40	4.00
		100	100	100.00	100.00	00.00	1.00	10.30
	G^{2*}	20	20	20.00	100.00	00.00	1.00	2.00
		40	40	40.00	100.00	00.00	1.00	3.80
		100	20	52.25	0.52	80.00	0.20	1.25
		100	40	76.56	0.77	60.00	0.40	3.89
		100	100	100.00	100.00	00.00	1.00	8.50
GRM	X^{2*}	20	20	20.00	100.00	00.00	1.00	4.44
		40	40	26.67	100.00	00.00	1.00	4.33
		100	20	15.33	0.15	80.00	0.20	4.67
		100	40	36.10	0.36	60.00	0.40	2.20
		100	100	52.00	100.00	00.00	1.00	7.30
	G^{2*}	20	20	20.00	100.00	00.00	1.00	4.10
		40	40	26.67	100.00	00.00	1.00	3.67
		100	20	23.60	0.24	80.00	0.20	3.60
		100	40	36.10	0.36	60.00	0.40	3.30
		100	100	52.00	100.00	00.00	1.00	6.00

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika* 43: 69-81.

Bjorner, J. B., Smith, K. J., Stone, C. A. & Sun, X. (2007). *IRTFIT: A macro for item fit and local dependence tests under IRT models*. Lincoln, RI: Quality Metric, Inc.

Ban, J. C., Hanson, B. A., Yi, Q. & Harris, D. J. (2002). *Data sparseness and online pretest item calibration/scaling methods in CAT.* ACT Research Report Series. Iowa City, IA: American College Testing Program.

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

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* 46 (4): 433-459.

Chon, K. H., Lee, W. & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement* 47: 318–338.

DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML. *Applied Measurement in Education* 15: 15-31.

Dodeen, H. (2004). The relationship between item parameters and item fit. *Journal of Educational Measurement* 41: 261–270.

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

Glas, C. A. W. (2000). Item calibration and parameter drift. In W.J. van der Linden & C. A. W. Glas (Eds.), *Computer adaptive testing: Theory and practice* (pp.183-200). Boston, MA: Kluwer-Nijhoff Publishing.

Glas, C. A. W. & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement* 68: 907-922.

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

Hambleton, R. K. & Han, N. (2004, April). *Assessing the fit of IRT models: Some approaches and graphical displays*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Harmes, J. C., Parshall, C. G. & Kromrey, J. K. (2003, April). *Recalibration of IRT item parameters in CAT: Sparse data matrices and missing data treatments.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Linacre, J. M. (2009). *Winsteps* (Version 3.69) [Computer Software]. Chicago, IL: Winsteps.com.

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

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−X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement* 27: 289–298.

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

McKinley, R. L. & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement* 9: 49–57.

Muthén, B., Asparouhov, T., Hunter, A. & Leuchter, A. (2011). Growth modeling with non-ignorable dropout: Alternative analyses of the STAR*D antidepressant trial. *Psychological Methods* 16: 17-33.

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

Muraki, E. & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating scale data* [computer program]. Chicago, IL: Scientific Software.

Rubin, D.B. (1976). Inference and missing data. *Biometrika* 63: 581-592.

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

SAS Institute Inc. (2008). *SAS/STAT® 9.2 user's guide*. Cary, NC: SAS Institute Inc.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology* 59: 429–449.

Sinharay, S. & Lu, Y. (2008). A further look at the correlation between item parameters and item fit statistics. *Journal of Educational Measurement* 45: 1–15.

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). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement* 40: 331–352.

Stone, C. A., Ankenmann, R. D., Lane, S. & Liu, M. (1993, April). *Scaling QUASAR's performance assessment.* Paper presented at the 1993 Annual Meeting of the American Educational Research Association, Atlanta, GA.

Stone, C. A., Mislevy, R. J. & Mazzeo, J. (1994). *Classification error and goodness-of-fit in IRT models*. Paper presented at the 1994 Annual Meeting of the American Educational Research Association, New Orleans, LA.

Thissen, D. & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer, N. Dorans, D. Eignor, R. Flaugher, B. Green, R. Mislevy, L. Steinberg & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (second edition). Hillsdale, NJ: Lawrence Erlbaum Associates, 101-133.

Thissen, D. & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika* 47: 397-412.

van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education* 13: 35-53.

Wang, S. & Harris, G. (2011). *Psychometric Evaluation of NWEA Item Calibration Procedure*. Technical Report. Portland, OR: Northwest Evaluation Association (NWEA).

Wang, S. & Jiao, H. (2012). *Examine Construct Validity of Computerized Adaptive Test in K-12 Assessments*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

Wells, C. S. & Bolt, D. M. (2004, April). *Investigation of a nonparametric procedure for assessing goodness-of-fit in item response theory*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

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