ACT Research Report Series 2007-1

An Investigation of the Performance of the Generalized S-X^2 Item-Fit Index for Polytomous IRT Models

Taehoon Kang

Troy T. Chen

ACT

June 2007

An Investigation of the Performance of the Generalized S-X^2 Item-Fit Index for Polytomous IRT Models

Taehoon Kang
Troy T. Chen

Abstract

Orlando and Thissen (2000, 2003) proposed an item-fit index, S-X^2, for dichotomous item response theory (IRT) models, which has performed better than traditional item-fit statistics such as Yen's (1981) Q_1 and McKinley and Mill's (1985) G^2. This study extends the utility of S-X^2 to polytomous IRT models, including the generalized partial credit model (GPCM: Muraki, 1992), partial credit model (PCM: Masters, 1982), and rating scale model (RSM: Andrich, 1978). The performance of the generalized S-X^2 in assessing item-model fit was studied in terms of empirical Type I error rates and power as compared to results obtained for G^2 provided by the computer program PARSCALE (Muraki & Bock, 1997). The results show that the generalized S-X^2 is a promising item-fit index for polytomous items in educational and psychological testing programs.

Keywords*: item response theory; item fit; S-X^2; polytomous IRT; PARSCALE*

An Investigation of the Performance of the Generalized S-X^2 Item-Fit Index for Polytomous IRT Models

Introduction

Ever since Lawley (1943) and Lord (1952) established the basic concepts of item response theory (IRT), many IRT models have been developed and applied to various fields of study such as psychological scaling and educational measurement. Unless an appropriate IRT model for a given data set is used, however, the benefits of IRT for applications such as test development, item banking, differential item functioning (DIF), computerized adaptive testing (CAT), and test equating might not be attained. In brief, the success of IRT applications requires satisfactory fit between the model and the data. The most critical problem caused by model-data misfit may be that the hallmark feature of IRT, parameter invariance, no longer applies (Shepard, Camilli & Williams, 1984; Bolt, 2002; Rupp & Zumbo, 2004).

Numerous statistical procedures have been developed to evaluate item fit under an IRT model, and goodness-of-fit studies have been conducted and reported in the voluminous IRT literature (see Bock, 1972; Douglas & Cohen, 2001; Glas & Suarez-Falcon, 2003; Liang & Wells, 2007; McKinley & Mills, 1985; Orlando & Thissen, 2000, 2003; Sinharay, 2003, 2005; Stone, 2000; Stone & Zhang, 2003; Suarez-Falcon & Glas, 2003; Wells, 2004; Yen, 1981). Among them, several Chi-square based item-level goodness-of-fit indices using significance tests such as Yen's (1981) Q_1 for dichotomous items, the traditional log-likelihood Chi-square, G^2, for both dichotomous and polytomous items (McKinley & Mills, 1985), and Orlando and Thissen's (2000, 2003) S-X^2 for dichotomous items have been utilized for IRT applications. Type I error rates for these goodness-of-fit indices have been investigated and reported.

A shortcoming of the item-fit tests based on Q_1 and G^2 is their sensitivity to test length and sample size. For instance, on a short test of 10 dichotomous items, these traditional statistics exhibited inflated empirical Type I error rates as high as 0.96 and 0.97, respectively, for a given nominal rejection rate of $\alpha = 0.05$ (Orlando & Thissen, 2000). DeMars' (2005) simulation studies with a sample size of 1,000 used PARSCALE's (Muraki & Bock, 1997) fit index which is similar to McKinley and Mill's (1985) G^2, and discovered empirical Type I error rates of 0.142 under the partial credit model (PCM: Masters, 1982) and 0.304 under the graded response model (GRM: Samejima, 1969) on a 10 polytomous item test. Stone and Hansen (2000) found that an item-fit test using G^2 for a 32 polytomous item test under the GRM showed inflated empirical Type I error rates between 0.142 and 0.181 for cases with 1,000 examinees, and between 0.229 and 0.396 for cases with 2,000 examinees, even though the true item parameter values were used in calculating predicted proportions.

Besides studies on Q_1 and G^2, there have been noteworthy studies on S-X^2 by Orlando and Thissen (2000, 2003). In their simulation studies using test lengths of 10, 40, and 80 dichotomous items and a sample size of 1,000, they showed that S-X^2 adequately controlled Type I error rates (Orlando & Thissen, 2000). Under the 1-, 2-, and 3-parameter logistic models (1PLM, 2PLM, and 3PLM, respectively), the empirical Type I error rates for tests based on S-X^2 were found to be between 0.04 and 0.07 with nominal α of 0.05. Additionally, the empirical power of S-X^2 improved as sample size increased from 500 to 2,000 (Orlando & Thissen, 2003).

The S-X^2 index could also be generalized and applied to the goodness-of-fit test for polytomous items (Roberts, in press). The main purpose of this study is to assess the performances of the generalized S-X^2 under the polytomous IRT models including the generalized partial credit model (GPCM: Muraki, 1992), PCM, and the rating scale model (RSM: Andrich, 1978) for different combinations of test length and sample size. The paper

begins with a review of Q_1, G^2, and S-X^2 statistics followed by a discussion on the generalization of S-X^2 for polytomous items. Finally, the performances of the generalized S-X^2 and PARSCALE's G^2 are compared through a simulation study.

Chi-Square Based Item Fit Indices

According to Hambleton, Swaminathan and Rogers (1991), and Stone (2000), a common strategy for assessing item-fit of an IRT model can be summarized as follows: (1) estimate the item and ability parameters under the chosen model, (2) classify examinees into *K* homogeneous groups in terms of their ability estimates, (3) calculate observed response proportions in each group for the item under investigation, (4) derive predicted response proportions in each group using the item and ability parameter estimates under the IRT model of interest, and (5) compute Chi-square based statistics by comparing the observed and predicted values.

The Traditional Chi-Square Based Fit Indices

Both Q_1 and G^2 are considered to be traditional Chi-square based fit indices. Yen's (1981) Q_1 for a dichotomous item *i* can be expressed as follows:

$$Q_{1i} = \sum_{k=1}^{10} \sum_{z=0}^{1} N_k \frac{(O_{ikz} - E_{ikz})^2}{E_{ikz}} = \sum_{k=1}^{10} N_k \frac{(O_{ik1} - E_{ik1})^2}{E_{ik1}(1 - E_{ik1})}, \qquad (1)$$

where *z* indicates the item score, *k* represents the number of groups of examinees, N_k is the number of examines in group *k*, and O_{ik1} (= 1-O_{ik0}) and E_{ik1} (= 1-E_{ik0}) are, respectively, the observed and predicted proportions of correct responses for group *k*. To compute Q_1, the item and ability parameter estimates are first obtained under the chosen IRT model, and the total number of groups (*K*) is set to 10 with each group having approximately an equal number of examinees. The predicted proportions of correct responses, E_{ik1}, is computed as the mean predicted proportion of each group. Since *K* = 10, the degrees of freedom (*df*)

associated with Q_1 equal 10 – *m* where *m* is the number of item parameters estimated.

For assessing model-item fit for both dichotomous and polytomous items, PARSCALE provides G^2 as the goodness-of-fit index. Given an item denoted *i*, G^2 can be computed as follows:

$$G_i^2 = 2\sum_{k=1}^{K_i}\sum_{z=0}^{Z_i} r_{ikz} \ln \frac{r_{ikz}}{N_{ik}P_{iz}(\bar{\theta}_k)}, \qquad (2)$$

where z indicates item scores ranging from zero to the highest item score of Z_i, *k* represents the number of groups of examinees, r_{ikz} equals the observed number of examinees scoring z in group k, N_{ik} is total number of examinees in group k, and $P_{iz}(\bar{\theta}_k)$ is the response function for item score *z* evaluated at the mean ability of examinees in group k.

The total number of groups K_i could however vary across items because neighboring groups can be collapsed to avoid expected values, $N_{ik}P_{iz}(\bar{\theta}_k)$, less than 5. In general, the *df* of G_i^2 for dichotomous items equals K_i which is different from that of Yen's Q_1 as no adjustment for the number of estimated parameters *m* is made for G_i^2. Mislevy and Bock (1990) argued that *m* is not considered in determining the *df* for G^2 because the parameter estimation procedure has nothing to do with minimizing G^2.

Orlando and Thissen's S-X^2 Index for Dichotomous Items

Though Orlando and Thissen's (2000) S-X^2 procedure follows much the same pattern as the Q_1 procedure, it has a notable advantage over Q_1 and G^2. For both Q_1 and G^2, the grouping procedure relies on sample- and model- dependent cut scores, whereas S-X^2 is based on test scores (i.e., number-correct scores).

Using the same notation defined earlier, S-X^2 for a dichotomous item *i* on an *I*-item test is given by:

$$\text{S-X}^2 = \sum_{k=1}^{I-1} \sum_{z=0}^{1} N_k \frac{(O_{ikz} - E_{ikz})^2}{E_{ikz}} = \sum_{k=1}^{I-1} N_k \frac{(O_{ik1} - E_{ik1})^2}{E_{ik1}(1 - E_{ik1})} \quad . \tag{3}$$

The summation over the test score k in equation (3) excludes $k = 0$ and $k = I$, since the proportion of examinees who respond to an item correctly and have a test score of zero is always zero, and it is always 1 for those having a perfect test score of I.

Also, in equation (3), the expected proportion of examinees in $x = k$ group who got item i right, $E_{ikI,}$ could be calculated using the following formula:

$$E_{ik1} = \frac{\int P_{i1}(\theta) f^{*i}(k-1 \mid \theta) \phi(\theta) \partial\theta}{\int f(k \mid \theta) \phi(\theta) \partial\theta}, \tag{4}$$

where $f(k \mid \theta)$ is the conditional predicted test score ($x = k$) distribution given θ, $f^{*i}(k \mid \theta)$ represents the conditional predicted test score distribution without item i, and $\phi(\theta)$ is the population distribution of θ. $f(k \mid \theta)$ and $f^{*i}(k \mid \theta)$ could be computed with the help of the recursive algorithm developed by Lord and Wingersky (1984).

Similar to G^2, neighboring groups could be collapsed to maintain a minimum expected cell frequency of 5. For S-X^2, the minimum value is set to 1 according to the studies of Larntz (1978) and Orlando and Thissen's (2000). If it is not necessary to collapse groups, then the *df* equals $I - 1 - m$ where m is the number of item parameters estimated; otherwise an adjustment for the number of groups being collapsed is needed.

The Generalized S-X^2 Index for Polytomous Items

The current study extends the application of S-X^2 to the assessment of item fit for polytomous items. For a polytomous item denoted i on a test of I polytomous items with each having $Z_i + 1$ scoring categories (i.e., category score $z = 0, 1, \ldots, Z_i$), the generalized S-X^2 can, using notation defined earlier, be expressed as follows:

$$\text{S-X}^2 = \sum_{k=Z_i}^{(F-Z_i)} \sum_{z=0}^{Z_i} N_k \frac{(O_{ikz} - E_{ikz})^2}{E_{ikz}}, \quad (5)$$

where z indicates category scores, Z_i is the highest score of item i, and F is a perfect test score (i.e., $F = \sum_i^I Z_i$). The expected category proportions, E_{ikz}, in equation (5), can be computed using the following formula:

$$E_{ikz} = \frac{\int P_i(z \mid \theta) f^{*i}(k - z \mid \theta) \phi(\theta) \partial\theta}{\int f(k \mid \theta) \phi(\theta) \partial\theta}. \quad (6)$$

To compute $f(z \mid \theta)$ and $f^{*i}(z \mid \theta)$ in equation (6), the generalized recursive algorithm developed by Thissen, Pommerich, Billeaud and Williams (1995) can be used.

Similar to that for dichotomous items, the computation for the generalized S-X^2 procedure excludes groups with a test score of zero (i.e., k = 0) or perfect test score (i.e., k = F). Also, in equation (5), the summation for k is from Z_i through $F - Z_i$. This is because within some groups with extremely low or high test scores, the expected proportions of examinees (E_{ikz}) for some categories are always zero. For instance, for the group with k = 38 on a test of 10 polytomous items with each having 5 categories (i.e., z = 0, 1, 2, 3, 4) in Table 1, the observed and predicted proportions of examinees for the z = 0 and z = 1 categories are always zero. Similarly, for the group with k = 2, they will always be zero for the z = 3 and z = 4 categories. For the generalized S-X^2 procedure, such groups are collapsed to the groups with $k = Z_i$ or $k = F - Z_i$. For example, the groups with k = 1, 2 and 3 for the illustrative item in Table 1 will be combined with the k = 4 group, and the groups with k = 37, 38, and 39 will be merged with the k = 36 group. Hence, for an item with $Z_i + 1$

scoring categories, the total number of groups included in the generalized S-X^2 procedure is equal to $K_i = F - 2Z_i + 1 = \left(\sum_i^I Z_i \right) - 2Z_i + 1$.

TABLE 1

The Observed and Expected Cell Frequencies of an Illustrative Item

Test Score Group *k*	Observed Frequencies ($N_k O_{ikz}$)					Expected Frequencies($N_k E_{ikz}$)				
	Item Category Score					Item Category Score				
	0	1	2	3	4	0	1	2	3	4
1	**0**	**0**	#	#	#	**0**	**0**	#	#	#
2	**4**	**0**	**0**	#	#	**3.34**	**.65**	**.01**	#	#
3	**6**	**1**	**0**	**0**	#	**5.20**	**1.76**	**.05**	*	#
4	4	1	0	0	0	3.24	1.68	**.07**	*	*
5	8	9	1	0	0	10.05	7.44	**.50**	**.01**	*
6	10	5	1	0	0	7.57	7.68	**.73**	**.02**	*
7	7	5	1	0	0	5.14	6.92	**.90**	**.03**	*
8	4	7	3	0	0	4.58	7.99	1.37	**.06**	*
9	6	15	1	0	0	5.87	13.07	2.88	**.18**	*
10	7	18	10	1	0	7.75	21.69	6.07	**.48**	*
11	3	20	5	0	0	4.80	16.73	5.87	**.59**	**.01**
12	2	30	7	0	0	5.26	22.63	9.86	1.23	**.02**
13	5	28	8	0	0	4.28	22.64	12.16	1.89	**.03**
14	3	19	18	3	0	3.41	22.12	14.60	2.80	**.06**
15	0	14	9	6	0	1.72	13.61	11.01	2.59	**.07**
16	3	16	18	5	0	1.82	17.60	17.39	5.02	**.17**
17	1	12	23	9	1	1.43	16.82	20.29	7.17	**.30**
18	1	15	17	10	1	**.95**	13.72	20.18	8.70	**.44**
19	1	14	21	15	0	**.76**	13.26	23.74	12.47	**.77**
20	2	6	19	9	1	**.37**	7.84	17.07	10.90	**.82**
21	0	9	20	14	3	**.30**	7.77	20.54	15.93	1.46
22	0	6	21	18	2	**.20**	6.19	19.87	18.68	2.07
23	1	3	15	18	2	**.10**	3.92	15.27	17.38	2.33
24	0	3	13	18	3	**.06**	2.78	13.13	18.09	2.93
25	0	2	19	20	2	**.04**	2.37	13.56	22.60	4.42
26	0	2	11	23	6	**.02**	1.66	11.55	23.27	5.50
27	0	1	6	21	5	**.01**	**.92**	7.75	18.92	5.40
28	0	1	9	13	6	**.01**	**.55**	5.71	16.90	5.82
29	0	0	3	19	3	*	**.32**	4.04	14.56	6.07
30	0	1	1	12	8	*	**.18**	2.86	12.58	6.38
31	0	0	2	16	10	*	**.15**	2.84	15.43	9.58
32	0	0	4	8	6	*	**.06**	1.38	9.37	7.19
33	0	0	0	11	8	*	**.03**	1.06	9.12	8.79
34	0	0	0	2	6	*	**.01**	**.31**	3.44	4.25
35	0	0	0	2	4	*	*	**.14**	2.22	3.63
36	0	0	0	1	5	*	*	**.08**	1.82	4.10
37	#	**0**	**0**	**1**	**7**	#	*	**.05**	**1.85**	**6.10**
38	#	#	**0**	**1**	**5**	#	#	**.01**	**.93**	**5.06**
39	#	#	#	**0**	**2**	#	#	#	**.16**	**1.84**

# indicates that the cell has a value of zero always.

* indicates that the expected value is less than 0.005 but larger than 0.

In light of Orlando and Thissen (2000), neighboring test score groups need to be collapsed to maintain a minimum expected category frequency ($N_k E_{ikz}$) of 1. However, this collapsing method could be infeasible for the case of polytomous items. For example, if Orlando and Thissen's cell collapsing approach is applied to the illustrative item in Table 1, only two groups with k = 18 and 19 would remain, and too much information would be lost. An even worse scenario would be that only one group remains.

Thus, instead of collapsing test score groups, the current study suggests, if needed, collapsing adjacent cells of item score categories for a given group k to ensure a minimum expected category frequency of 1. For instance, in the group with k = 11 of the illustrative item in Table 1, the z = 3 and z = 4 categories could be combined with the z = 2 category. Pett (1997) mentioned that cell-collapsing should be undertaken carefully so that the combined cells make intuitive sense. Since the item score categories are ordered, the suggested method is considered reasonable. Also, Muraki (1996) used this approach for combining response categories. With this collapsing algorithm, the generalized S-X^2 has *df* of $K_i Z_i - m - C_i$ where m is the number of item parameters estimated, and C_i indicates the total number of item score categories being collapsed. For the illustrative item with 5 scoring categories in Table 1, the *df* = $33 * 4 - 5 - 61 = 66$ under the GPCM. Also, the generalized S-X^2 for this item was found to be 62.11 (p-value = 0.61). So, the GPCM appeared to fit the item adequately.

Method

Design of Simulation Study

To assess the performance of the generalized S-X^2 index, a simulation study varying polytomous IRT models, test lengths, and sample sizes was conducted. The index selected for comparison was PARSCALE's G^2 because it appears to be the most frequently employed index in applied settings. Three commonly used polytomous IRT models (RSM,

PCM and GPCM) were explored in this study. Under the GPCM, the probability that an examinee *j* scores *z* with *z* = 0, 1, …, Z_i on item *i* with $Z_i + 1$ response categories is modeled by the following:

$$P(z \mid \theta_j, \alpha_i, \beta_i, \tau_{ci}) = \frac{exp \sum_{c=0}^{z} \alpha_i \left[\theta_j - (\beta_i - \tau_{ci})\right]}{\sum_{y=0}^{Z_i} exp \sum_{c=0}^{y} \alpha_i \left[\theta_j - (\beta_i - \tau_{ci})\right]}, \quad (7)$$

where α_i is the discrimination of item *i*, β_i denotes the difficulty of item *i*, and τ_{ci} represents the location parameter for a category on item *i*. Equation (7) needs to set $\tau_{0i} = 0$, $\sum_{c=1}^{Z_i} \tau_{ci} = 0$ and $exp \sum_{c=0}^{0} \alpha_i \left[\theta_j - (\beta_i - \tau_{ci})\right] = 1$ for model identification.

The three polytomous IRT models considered in this study are hierarchically related. If α_i in equation (7) is fixed at 1 across items, equation (7) reduces to the PCM. Moreover, if the τ values for each category are, respectively, the same across items, equation (7) further reduces to the RSM. Consequently, RSM is nested within PCM, while PCM is nested within GPCM.

In addition to the three generating models (RSM, PCM, and GPCM), this simulation study employed three test lengths (*I* = 5, 10, and 20 items), and four sample sizes (*N* = 500, 1,000, 2,000, and 5,000 examinees). The three test lengths mimic tests having small, moderate and large numbers of polytomously scored items. The four sample sizes represent small, moderate, large and very large samples.

Data Generation

A standard method was employed for item response generation for this study. The steps for data generation include: (1) generate item and ability parameters, (2) under the chosen IRT model, calculate the probability, $P(z \mid \theta_j, \alpha_i, \beta_i, \tau_{ci})$, for the responses using the

generated item and ability parameters, and then the cumulative probability denoted $P^*(z \mid \theta_j, \alpha_i, \beta_i, \tau_{ci}) = \sum_{x=0}^{z} P(x \mid \theta_j, \alpha_i, \beta_i, \tau_{ci})$, and (3) with a random number denoted u drawn from the uniform distribution, *U(0,1)*, assign a response of 0 if $u \le P^*(0)$, or z if $P^*(z-1) < u \le P^*(z)$ for z = 1, 2, …, or Z_i.

For this study, the item parameters used for simulating response data under the GPCM were obtained as follows. The discrimination parameters (α_i) were randomly sampled from a lognormal (0, $.5^2$) distribution. For each item, four item step parameters (i.e., $\delta_{ci} = \beta_i - \tau_{ci}$ where c = 1, 2, 3, or 4) were randomly drawn from four normal distributions with a common standard deviation of 0.5 and means of -1.5, -0.5, 0.5, and 1.5, respectively. The mean of these four step parameters is then used as the item difficulty parameter (β_i), and the difference between β_i and δ_{ci} is taken as τ_{ci}. This item-parameter generating procedure was repeated for all I items. The values for the θ parameter were randomly drawn from the standard normal distribution, *N(0,1)*. With these item and ability parameters, a response dataset under the GPCM was generated. The generating procedure for datasets under the PCM is the same as that for the GPCM except that the α_i parameters were fixed at 1. For generating the dataset under the RSM, only one random sample of each step parameters (τ_{ci}) was generated and used for all items, while the discrimination parameters for each item were also fixed at 1.

Finally, there were a total of 36 different conditions simulated in this study (3 generating models × 3 test lengths × 4 sample sizes). One hundred replications were generated for each condition, and each condition mimicked 100 different I-item tests from the same item pool administered to 100 equivalent groups of n examinees, respectively.

Item Parameter Calibration

The item and ability parameters in each simulated dataset were calibrated using the computer program PARSCALE (Muraki & Bock, 1997). Examples of the PARSCALE codes used for GPCM, PCM, and RSM calibration are given in Appendix A. Paralleling Orlando and Thissen's (2000) study, the number of item parameters of the calibration model (CM) was always fewer than or equal to that of the generating model (GM). As shown in Table 2, when the CM and GM were the same, results were used to evaluate the empirical Type I error rates of the generalized S-X^2 and PARSCALE's G^2. And, when the CM was a simpler model than the GM, results were used to calculate the empirical power of the item-fit indices.

TABLE 2

Model Calibration Design in the Simulation Study

Calibrating Model (CM)	Generating Model (GM)		
	RSM	PCM	GPCM
RSM	Type I Error Rate	Power	Power
PCM	-	Type I Error Rate	Power
GPCM	-	-	Type I Error Rate

This study used a nominal α of 0.05. An item was flagged for misfit if the significance level (i.e., p-value) for the observed fit index under investigation was less than 0.05. Under each condition, the simulated Type I error rates or power for the fit indices were obtained by dividing the number of flagged items by $100 \times I$ upon the completion of the 100 replications.

Results

When the CM was same as the GM, the PARSCALE calibrations ran properly and converged successfully for all study conditions. However, when the CM is a simpler model than the GM, the calibration runs did not converge successfully for some conditions with 100 EM cycles. The worst situation occurred for the condition with 20 items and 2,000 examinees under CM = PCM and GM = GPCM. For this condition, 24 out of 100 data sets were not successfully calibrated. In the current study, the problematic cases were excluded from the calculations of empirical power.

The proportions of items wrongly flagged for misfit are shown in Table 3 for each condition. At a glance, the simulated Type I error rates of the generalized S-X^2 did not change much across conditions, while the values for the PARSCALE's G^2 differed drastically. In the short (5-item) and medium (10-item) test length conditions, the Type I error rates of G^2 appeared to be severely inflated in most cases. This is not surprising because it is a well-known problem for the traditional item-fit indices as mentioned earlier. Also, even though the test length is as long as 20 items, G^2 had inflated Type I error rates in the 5,000 examinee conditions (0.213 for the RSM, 0.118 for the PCM, and 0.074 for the GPCM). In contrast to the poor performance of G^2, the generalized S-X^2 did not seem to be affected by test length, and had empirical Type I error rates ranging from 0.040 and 0.065 for all study conditions except those with long (20-item) test length and small sample size of 500 examinees. In the long (20-item) test conditions, the performance of S-X^2 appeared to improve consistently as the sample size increased from 500 to 2,000, but this pattern was not observed for G^2.

TABLE 3

Type I Error Rates: Proportions of Indices with Significance Level Greater Than .05, Under GM = CM

Test Length and Sample Size		Generalized S-X^2			PARSCALE G^2		
		RSM	PCM	GPCM	RSM	PCM	GPCM
5	500	0.048	0.050	0.058	0.432	0.310	0.350
	1,000	0.044	0.048	0.046	0.816	0.762	0.676
	2,000	0.050	0.052	0.052	0.998	1.000	0.930
	5,000	0.064	0.060	0.040	1.000	1.000	0.994
10	500	0.048	0.048	0.065	0.085	0.048	0.034
	1,000	0.063	0.056	0.058	0.177	0.103	0.100
	2,000	0.061	0.056	0.042	0.424	0.280	0.290
	5,000	0.047	0.046	0.052	0.926	0.875	0.654
20	500	0.095	0.098	0.126	0.055	0.024	0.015
	1,000	0.065	0.064	0.054	0.061	0.027	0.015
	2,000	0.057	0.054	0.049	0.088	0.051	0.025
	5,000	0.046	0.055	0.053	0.213	0.118	0.074

Table 4 shows the empirical power of the generalized S-X^2 and G^2 when the GM is more complex than the CM. As expected, the greater the difference that exists between GM and CM, the higher the power is. When CM was RSM, the generalized S-X^2 was found to be more sensitive in detecting misfit with GM = GPCM than with GM = PCM. Similarly, for the conditions with GM = GPCM, the conditions with CM = RSM yield higher power than those with CM = PCM. For all combinations of test length and sample size, the highest power was found when GM = GPCM and CM = RSM. These findings reflect the nested structure of the three IRT models. Although both the generalized S-X^2 and G^2 showed better power as the sample size increased, the power values for G^2 were not very useful because of the inflated Type I error rates of this index under most study conditions.

TABLE 4

Empirical Power: Proportions of Indices with Significance Level Greater Than .05, Under GM > CM

Test Length and Sample Size		Generalized S-X^2			PARSCALE G^2		
		PCM > RSM	GPCM > RSM	GPCM > PCM	PCM > RSM	GPCM > RSM	GPCM > PCM
5	500	0.368	0.615	0.252	0.756	0.731	0.389
	1,000	0.606	0.829	0.415	0.966	0.896	0.672
	2,000	0.808	0.941	0.625	1.000	0.982	0.848
	5,000	0.942	0.981	0.755	1.000	0.998	0.979
10	500	0.289	0.565	0.293	0.562	0.755	0.432
	1,000	0.536	0.793	0.459	0.802	0.889	0.565
	2,000	0.734	0.918	0.609	0.949	0.973	0.754
	5,000	0.929	0.980	0.733	0.998	0993	0.877
20	500	0.302	0.571	0.341	0.494	0.805	0.480
	1,000	0.471	0.740	0.406	0.751	0.926	0.630
	2,000	0.673	0.888	0.550	0.897	0.970	0.755
	5,000	0.889	0.973	0.714	0.971	0.996	0.896

Discussion

As the empirical Type I error rates of G^2 were found to be inappropriately inflated, its performance is not discussed further in this section. The discussion presented here focuses on issues related to the generalized S-X^2 index. This section begins with a scrutiny on the *df* adjustment for the number of estimated item parameters followed by the discussion on the performance of the generalized S-X^2.

The df Adjustment for the Number of Estimated Item Parameters

For the applications of item-fit indices, controversies over the *df* adjustment for the number of estimated item parameters have been found in the IRT literature. For instance, the *df* of Yen's Q_1 is adjusted for the number of estimated item parameters, while the *df* of PARSCALE's G^2 is not adjusted. DeMars (2005) explained that the disagreement comes from the difference in item parameter estimation methods. Yen's Q_1 is designed for item-fit analysis when the joint maximum likelihood (JML) method is used. In contrast, PARSCALE employs the marginal maximum likelihood (MML). Mislevy and Bock (1990) mentioned, for the MML approach, "the residuals are not under linear constraints and there is

no loss of degrees of freedom due to the fitting of the item parameters (p.1-11)."

Even though Orlando and Thissen (2000, 2003) used the MML method with the computer program MULTILOG (Thissen, 1991), they applied the *df* adjustment. An argument for using the *df* adjustment was given by Stone and Zhang (2003): Regardless of item parameter estimation approach, the *df* adjustment should be used "to account for the fact that expected frequencies are based on estimated item parameters (p. 347)." Applying the *df* adjustment to their studies on S-X^2 for dichotomous items, both Orlando and Thissen (2000) and Stone and Zhang (2003) found empirical Type I error rates to be close to the nominal rates.

Based upon these discussions, a preliminary investigation was conducted to compare the applications of the generalized S-X^2 *with* and *without* the *df* adjustment. The results showed no dramatic difference in the simulated Type I error rates and power even though the cases without the *df* adjustment were slightly more conservative. Additionally, since the generalized S-X^2 was derived on the basis of Orlando and Thissen's S-X^2, the *df* adjustment was utilized for the current study.

Performance of the Generalized S-X^2

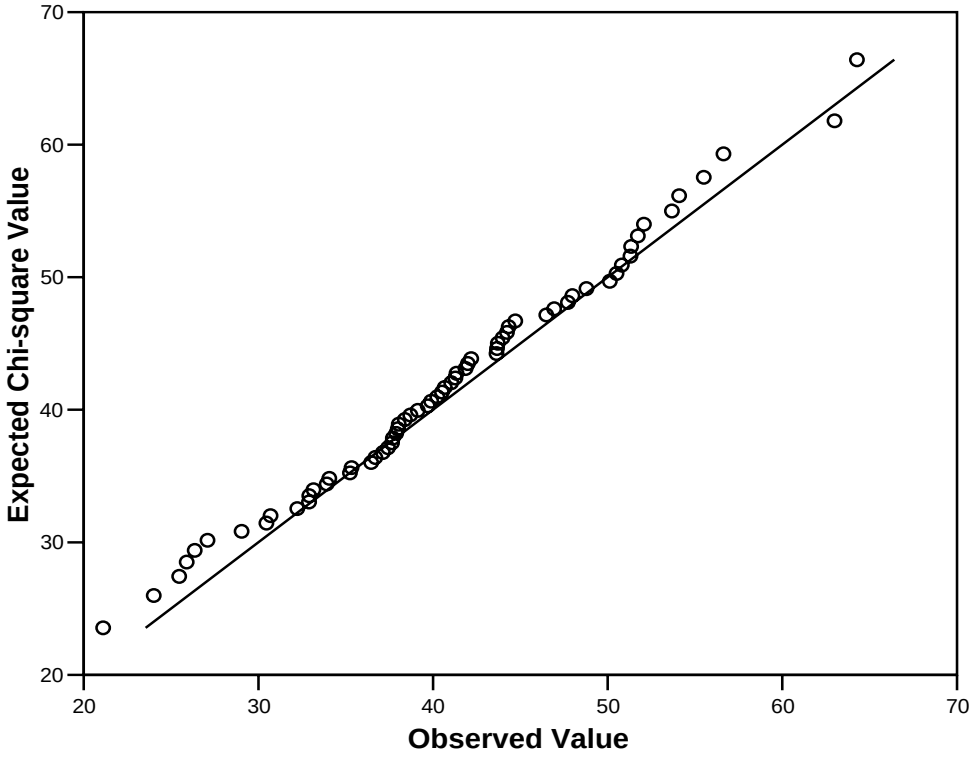
Similar to Orlando and Thissen's (2000) study on S-X^2 for dichotomous items, the current study found that the generalized S-X^2 exhibited empirical Type I error rates ranging from 0.040 to 0.065 for all study conditions except that with test length of 20 items and number of examinees equal to 500. For this condition, the Type I error rates of the generalized S-X^2 appeared to be inflated as high as 0.095 for the RSM, 0.098 for the PCM, and 0.126 the GPCM, as shown in Table 3. This could be explained by the inevitable sparseness in expected frequencies. Given that there are more score categories and hence more total score groups on an *I* polytomous item test than on an *I* dichotomous item test, the applications of generalized S-X^2 would encounter more sparseness in expected frequencies

for conditions with long tests (e.g., 20 items) and small sample sizes (e.g., 500).

Another issue related to control of Type I errors is how close the sampling distribution of the item-fit statistics under the null hypothesis is to the theoretical distribution. To study the extent to which the null distribution of the S-X^2 index followed a Chi-square distribution, Orlando and Thissen (2000) examined the first two moments of the index. For a Chi-square distribution, the mean and variance equal the *df* and $2 \times df$, respectively. Also, Stone (2000) used the technique of Q-Q plots to compare the empirical distribution of the fit statistics of G^2 to the expected Chi-square distribution with an estimated *df*. These two methods were employed for examining the empirical distribution of the generalized S-X^2 index.

It is believed that the generalized S-X^2 has approximately a Chi-square distribution. However, for a given test length and a chosen IRT model, the *df* would vary due to the total number of cells being collapsed. For example, the observed *df* ranged from 25 to 47 for the 100 replicates under the condition with test length of 5 items and sample size of 5,000 in the Type I error rate analysis under the GPCM. From the 500 simulated items (5 items $\times$ 100 replicates) under this condition, the observed mode of the *df* was 42 with a frequency count of 65. The mean and variance of these 65 empirical S-X^2 values were found to be 41.00 and 82.99, respectively. For other conditions, the two moments were also found to be close to their expected values. In addition, Figure 1 presents the Q-Q plot to compare the empirical distribution of the 65 S-X^2 observations to a theoretical Chi-square distribution with $df = 42$. The plot has a slope and intercept close to 1 and 0, respectively.

FIGURE 1. Q-Q Plot of Empirical S-X^2 Distribution Compared to a Chi-Square Distribution with df = 42



The results of power analysis in a simulation study for item-fit statistics are closely related to how to generate misfitting items under the non-null cases. As shown in Table 4, the more different the GM and CM were, the better the observed power analysis results. Therefore, the generalized S-X^2 had the highest power when the model fit of the RSM was investigated for the data simulated under the GPCM. Also, for these conditions, a moderate sample size of 1,000 would yield adequate power. When the GPCM was the GM and the PCM was the CM, however, a very large sample size (e.g., 5,000 examinees) was required to yield acceptable power higher than 0.7 regardless of the test length. Similarly, when the PCM was the GM and the RSM was the CM, sample sizes of 2,000 or higher are needed for the generalized S-X^2 index to produce satisfactory power.

Conclusion

Similar to the findings reported in the literature, the results of the current study showed that the performance of G^2 appeared to be poor in most conditions with short and moderate test lengths or very large sample sizes (e.g., 5,000 examinees). In contrast, for such tests with 5 or 10 polytomous items, the generalized S-X^2 showed superior performance in terms of Type I error rate control and power. Similar performance pattern of the index was also found for the cases with very large sample sizes regardless of test length. Consequently, the generalized S-X^2 is a promising index for investigating item-fit in educational and psychological assessments.

To gain a better understanding of this promising item-fit index, additional studies need to be conducted. First, with the design of this simulation study, all generated items were considered misfit for the power study. Following Orlando and Thissen's (2003) study, the sensitivity of the generalized S-X^2 in detecting different percentages of misfit items on a test form needs to be further studied. Second, the number of test score groups and hence the *df* are determined by the test length and the numbers of item score categories. All of the simulated items in the current study had a fixed number of score categories. Thus, studies on the impact due to different numbers of item score categories could provide some insight on the behavior of the index. Third, it is not rare for ability distributions to be non-normal for an educational or psychological assessment (Micceri, 1989). The performance of the generalized S-X^2 needs to be investigated under conditions where the ability distribution is not normal (e.g., uniform or skewed). Finally, noteworthy behavior of S-X^2 has been found for dichotomous and polytomous items separately, yet it might behave differently for mixed format tests because of format effect or multidimensionality. So, further studies on the S-X^2 index for mixed format test data are required.

References

Andrich, D. (1978). Application of a psychometric model to ordered categories which are scored with successive integers. Applied Psychological Measurement, 2, 581-594.

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

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. Applied Measurement in Education, 15, 113-141.

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

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., & 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., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA:Sage.

Larntz, K. (1978). Small-sample comparisons of exact levels for Chi-squared goodness-of-fit statistics. Journal of the American Statistical Association, 73, 253-263.

Lawley, D.N. (1943). On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 61A, 273-287.

Liang, T., & Wells, C. S. (2007). A model fit statistic for generalized partial credit model. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Lord, F.M. (1952). A theory of test scores. Psychometric Monograph, No. 7.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equationgs". Applied Psychological Measurement, 8, 453-461.

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.

Micceri, T. (1989). The unicorn, the normal curve and other improbable creatures. Psychological Bulletin, 105:1, 156-166.

Mislevy, R. J., & Bock, R. D. (1990). BILOG: Item analysis and test scoring with binary logistic models [computer software]. Chicago, IL: Scientific Software.

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

Muraki, E. (1996). A generalized partial credit model. In Van der Linden, W. J., & Hambleton, R. K. (Eds.), Handbook of modern item response theory. New York: Springer.

Muraki, E. & Bock, R. D. (1997). PARSCALE: IRT item analysis and test scoring for rating-scale data [Computer software]. Chicago: Scientific Software.

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.

Pett, M. A. (1997). Nonparametric statistics for health care research: Statistics for small samples and unusual distributions. Thousand Oaks, CA: Sage Publication, Inc.

Roberts, J. S. (in press). Modified likelihood-based item fit statistics for the generalized graded unfolding model. Applied Psychological Measurement.

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When Pearson correlations are not enough. Educational and Psychological Measurement, 65, 588-599.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, 17.

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Sinharay, S. (2003). Bayesian item analysis for dichotomous item response theory models (ETS RR-03-34). Retrieved May 10, 2005, from the ETS Web site http://www.ets.org/research/researcher/rr0334.htm.

Sinharay, S. (2005). Bayesian item fit analysis for unidimensional item response models. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

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., & Hansen, M. A. (2000). The effect of errors in estimating ability on goodness-of-fit tests for IRT models. Educational and Psychological Measurement, 60, 974-991.

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.

Suarez-Falcon, J. C., & Glas, C. A. W. (2003). Evaluation of global testing procedures for item fit to the Rasch model. British Journal of Mathematical and Statistical Psychology, 56. 127-143.

Thissen, D. (1991). MULTILOG: Multiple, categorical item analysis and test scoring using item response theory. Chicago, IL: Scientific Software. [Computer Program.]

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.

Wells, C. S. (2004). Investigation of model misfit in IRT and a new approach based on simultaneous parametric and nonparametric IRT estimation. Unpublished doctoral dissertation. University of Wisconsin-Madison.

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

Appendix A

Examples of the PARSCALE Codes Used for GPCM, PCM and RSM Calibration

< A-1: GPCM >

```
---------------------------------------------------------------------------
>FILE DFNAME='ca001.dat', SAVE;
>SAVE PARM='gpgp22ca001.par';
>INPUT NIDW=3, NTOTAL=10, NTEST=1, LENGTH(10), NFMT=1;
(3A1,T1,10A1)
>TEST1 TNAME=gp22ca001, ITEM=(1(1)10), NBLOCK=10;
>BLOCK BNAME=SBLOCK1, NITEMS=1, NCAT=5, SCORING=(1,2,3,4,5), REPEAT = 10;
>CALIB PARTIAL, LOGISTIC, CYCLES=(100,1,1,1), SCALE=1.0, NQPTS=40, ITEMFIT=10;
>SCORE NOSCORE;
---------------------------------------------------------------------------
```

< A-2: PCM >

```
---------------------------------------------------------------------------
>FILE DFNAME='ca001.dat', SAVE;
>SAVE PARM='pgp22ca001.par';
>INPUT NIDW=3, NTOTAL=10, NTEST=1, LENGTH(10), NFMT=1;
(3A1,T1,10A1)
>TEST1 TNAME=gp22ca001, ITEM=(1(1)10), NBLOCK=10;
>BLOCK BNAME=SBLOCK1, NITEMS=1, NCAT=5, SCORING=(1,2,3,4,5), REPEAT = 10;
>CALIB PARTIAL, LOGISTIC, CYCLES=(100,1,1,1), SCALE=1.0, NQPTS=40,
ITEMFIT=10,SPRIOR, PRIORREAD;
>PRIORS SMU=(1(0)10), SSIGMA=(0.0000001(0)10);
>SCORE NOSCORE;
---------------------------------------------------------------------------
```

< A-3: RSM >

```
---------------------------------------------------------------------------
>FILE DFNAME='ca001.dat', SAVE;
>SAVE PARM='rgp22ca001.par';
>INPUT NIDW=3, NTOTAL=10, NTEST=1, LENGTH(10), NFMT=1;
(3A1,T1,10A1)
>TEST1 TNAME=gp22ca001, ITEM=(1(1)10), NBLOCK=1;
>BLOCK BNAME=SBLOCK1, NITEMS=10, NCAT=5, SCORING=(1,2,3,4,5);
>CALIB PARTIAL, LOGISTIC, CYCLES=(100,1,1,1), SCALE=1.0, NQPTS=40,
ITEMFIT=10,SPRIOR, PRIORREAD;
>PRIORS SMU=(1(0)10), SSIGMA=(0.0000001(0)10);
>SCORE NOSCORE;
---------------------------------------------------------------------------
```