Asia Pacific Educ. Rev. (2011) 12:89–96
DOI 10.1007/s12564-010-9082-4

Performance of the generalized S-X^2 item fit index for the graded response model

Taehoon Kang · Troy T. Chen

Received: 6 January 2009 / Revised: 8 February 2010 / Accepted: 9 February 2010 / Published online: 18 March 2010

Abstract The utility of Orlando and Thissen's (2000, 2003) S-X^2 fit index was extended to the model-fit analysis of the graded response model (GRM). The performance of a modified S-X^2 in assessing item-fit of the GRM was investigated in light of empirical Type I error rates and power with a simulation study having various conditions typically encountered in applied testing situations. The results show that the Type I error rates were controlled adequately around the nominal alpha by S-X^2. The power of the S-X^2 statistic was much lower when the source of misfit was multidimensionality than when it was due to discrepancy from the true GRM curves. Once the data size increased sufficiently, however, appropriate power was obtained regardless of the source of the item-misfit. In summary, the generalized S-X^2 appears to be a promising index for investigating item fit for polytomous items in educational and psychological assessments.

Keywords Item response theory · Item fit · S-X^2 · Graded response model

Introduction

Item response theory (IRT) employs a family of mathematical models designed to describe the performance of examinees on test items. Satisfactory model-data fit is critical if the benefits of IRT applications such as test development, item banking, differential item functioning (DIF), computerized adaptive testing (CAT), and test equating are to be attained. Although there is now extensive research literature on IRT, relatively less research has been done to help practitioners evaluate the model-item fit for polytomous item response data.

The S-X^2 item-fit statistic was originally proposed for dichotomous IRT models and was found to perform better than the traditional item-fit statistics such as Yen's (1981) Q_1 and McKinley and Mills (1985) G^2. Recently, Kang and Chen (2008) successfully generalized the use of S-X^2 to polytomous IRT models such as the generalized partial credit model (GPCM; Muraki 1992), the partial credit model (PCM; Masters 1982), and the rating scale model (RSM; Andrich 1978). According to the categorization of IRT models proposed by Thissen and Steinberg (1986), these models are the "divide-by-total" models, while the GRM belongs to the "difference models" category.

For the GRM (Samejima 1969), Stone and Hansen (2000) found that an item-fit test using the Pearson χ^2 statistics or a log-likelihood ratio index, G^2, suffered from inflated Type I error rates. Also DeMars' (2005) simulation studies used PARSCALE's (Muraki and Bock 1997) item fit index and discovered inflated empirical Type I error rates on a 10 polytomous item test under the GRM. In both aforementioned studies, the empirical power of the item-fit indices was not investigated.

As polytomously scored items have become increasingly popular for many psychological and educational testing programs, it is necessary to develop an item-fit index for polytomous items that adequately controls Type I error rates and shows appropriate power in detecting misfit items. This includes "difference models" such as the GRM. To answer this call, the current study investigated

T. Kang (✉)
Department of Education, Sungshin Women's University,
249-1, Dongseon-Dong 3-Ga, Seongbuk-Gu,
Seoul 136-742, Korea
e-mail: taehoonkang@gmail.com

T. T. Chen
ACT Inc., Iowa City, USA

the performance of the S-X^2 item-fit statistic (Orlando and Thissen 2000, 2003) under the GRM. Specifically, this paper begins with reviewing the generalized S-X^2 procedure for polytomous items. Then, the design for the simulation study investigating the performance of the generalized S-X^2 under the GRM in terms of empirical Type I error rates and power is discussed. Third, the results of the simulation are explained and summarized. Finally, the overall effectiveness of the S-X^2 statistic is addressed, and future research is suggested.

The generalized S-X^2 for polytomous items

As a modified item-fit index of Orlando and Thissen's (2000, 2003) S-X^2 that is only for dichotomous items, the generalized S-X^2 for polytomous items can be expressed as follows:

$$\mathrm{S}-\mathrm{X}^2=\sum_{k=Z_i}^{(F-Z_i)}\sum_{z=0}^{Z_i}N_k\frac{(O_{ikz}-E_{ikz})^2}{E_{ikz}}, \tag{1}$$

where Z_i is the highest score of a polytomous item i, z indicates each category score, and F is a perfect test score (i.e., $F=\sum_{i=1}^{I}Z_i$) when the total number of items in a test is I. k represents a homogeneous group of examinees, N_k is the number of examines in group k, and O_{ikz} and E_{ikz} are, respectively, the observed and predicted proportions of the z category response in item i for group k.

The expected category proportions, E_{ikz}, in Eq. 1, can be computed using the following formula

$$E_{ikz}=\frac{\int P_i(z|\theta)f^{*i}(k-z|\theta)\varphi(\theta)\partial\theta}{\int f(k|\theta)\varphi(\theta)\partial\theta}, \tag{2}$$

where $P_i(z|\theta)$ is the calculated probability that a person with θ gets an item score z on item i under the GRM, $f(\cdot|\theta)$ is the conditional predicted test score distribution given θ, $f^{*i}(\cdot|\theta)$ represents the conditional predicted test score distribution without item i, and $\varphi(\theta)$ is the population distribution of θ. To compute $f(\cdot|\theta)$ and $f^{*i}(\cdot|\theta)$ used in Eq. 2, the generalized recursive algorithm developed by Thissen et al. (1995) is used.

As shown in Eq. 1, the summation for k is from the highest score of item i, Z_i, through $F-Z_i$ which is the difference between the perfect test score and the highest item i score. This is because for some groups with extremely low or high test scores, the expected proportions of examinees (E_{ikz}) for some score categories are always zero. For example, suppose an achievement test has 10 polytomous items with each item having five categories ($z=0$, 1, 2, 3, and 4). The possible test scores range between 0 and 40. Obviously, for the group of $k=3$, the E_{ik4} will be always zero because the item score cannot be four when the total test score is three. Similarly, for the group of $k=37$, the E_{ik0} will be always zero. Therefore, valid groups for computing S-X^2 in this example include those having $k=4$ (Z_i) to $k=36$ ($F-Z_i$). The respondents in the groups with extremely low (i.e., $k=0$, 1, 2, and 3) and high (i.e., $k=37$, 38, 39, and 40) test scores are collapsed with the $k=4$ and $k=36$ groups, respectively.

When respondents belong to low (high) test score groups, they tend not to have high (low) item scores for a specific polytomous item. Consequently, it happens often that E_{ikz} becomes very small in some groups. To ensure a minimum expected cell frequency of 1, adjacent cells of item score categories for a given group k are to be collapsed. With this collapsing algorithm, the df of the modified S-X^2 in Eq. 1 is computed as $(F-2Z_i+1)\times Z_i-m-C_i$ where m is the number of item parameters and C_i indicates the total number of item score category cells being collapsed.

Method

Type I error rate study: design of simulation study and data generation

To assess the performance of the generalized S-X^2 index under the GRM, a simulation study varying in test lengths, sample sizes, number of categories per item, and ability distributions was conducted. The simulation study employed three test lengths ($I=5$, 10, and 20 items), three sample sizes ($N=500$, 1,000, and 2,000 examinees), two numbers of categories ($nc=3$ and 5), and two ability distributions (normal and uniform). The three test lengths mimicked educational or psychological testing programs having various numbers of polytomously scored items, and the three sample sizes represented small, moderate, and large samples. The different numbers of item score categories were considered to be practical in real world settings. Finally, the employment of a non-normal ability distribution was based upon Micceri's (1989) findings that it is not rare for trait or ability distributions to be non-normal. In sum, there were a total of 36 different conditions simulated in this Type I error rate study (3 test lengths $\times$ 3 sample sizes $\times$ 2 numbers of item score categories $\times$ 2 ability distributions). 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.

For the empirical Type I error rate study, the item parameters used for simulating response data under the GRM were obtained as follows. The discrimination parameters (α_i) were randomly sampled from a *Lognormal* $(0, 0.5^2)$ distribution. For each item, the threshold parameters (i.e., β_1,

…, β_{nc-1}) were randomly drawn from uniform distributions. For the conditions with $nc = 3$, the first threshold parameter (β_1) for each item was drawn from a uniform distribution, $U(-2,1)$, and the parameters for a successive category (β_2) were obtained by adding a random value from $U(.4,1)$ to β_1. A similar item parameter generating approaches were applied for the conditions with $nc = 5$: $U(-2,0)$ was used for generating β_1, and three random values from $U(.4,.7)$ were added cumulatively to obtain β_2, β_3, and β_4. The values for the θ parameter were randomly drawn from the standard normal distribution, $N(0,1)$ or $U(-3,3)$. Under each condition, the simulated Type I error rates were obtained by dividing the number of wrongly flagged items by 100 × number of simulated items in each data set.

Power study: design of simulation study and data generation

For an empirical power study of IRT model-item fit, it is necessary to generate atypical (i.e., bad or misfitting) items deviating from the expectations of the logistic model under investigation. According to Stone (2000) and Lee et al. (2002), two main causes of lack of fit between observed data and model predictions are (1) violations to the model assumptions (e.g., unidimensionality) and (2) inadequacies in estimation procedures (i.e., a huge discrepancy between the expected item characteristic curves and empirical curves as a result of small sample size, random error, or true non-logistic function). Following these research findings, this power study considered two types of bad items: One was *M*-type misfit due to multidimensionality, and the other was *D*-type misfit due to a curve discrepancy. To generate these two different types of misfit items, a two-dimensional GRM (for *M*-type) and a Guttman step function (for *D*-type) were utilized.

When an item in a test measures a totally different construct (θ_2) from that (θ_1) measured by all the other items [i.e., $\rho(\theta_1, \theta_2) = 0$], this item is considered as a problematic item under the IRT assumption of unidimensionality. This *M*-type misfit can be explained with Fig. 1. As shown in Fig. 1a and c, the illustrative *M*-type polytomous item ($z = 0$, 1, 2, 3, and 4) is very useful in measuring θ_2. But, when the other items in the same test mainly measure θ_1, the item would appear to be not-discriminating under the unidimensional calibration model as shown in Fig. 1b. Consequently, it is anticipated that the data points could be explained with zero discrimination parameter or could not be explained effectively under the application of a unidimensional IRT model.

To generate a dataset including a single *M*-type misfit item and GRM-fitting items, the two-dimensional GRM

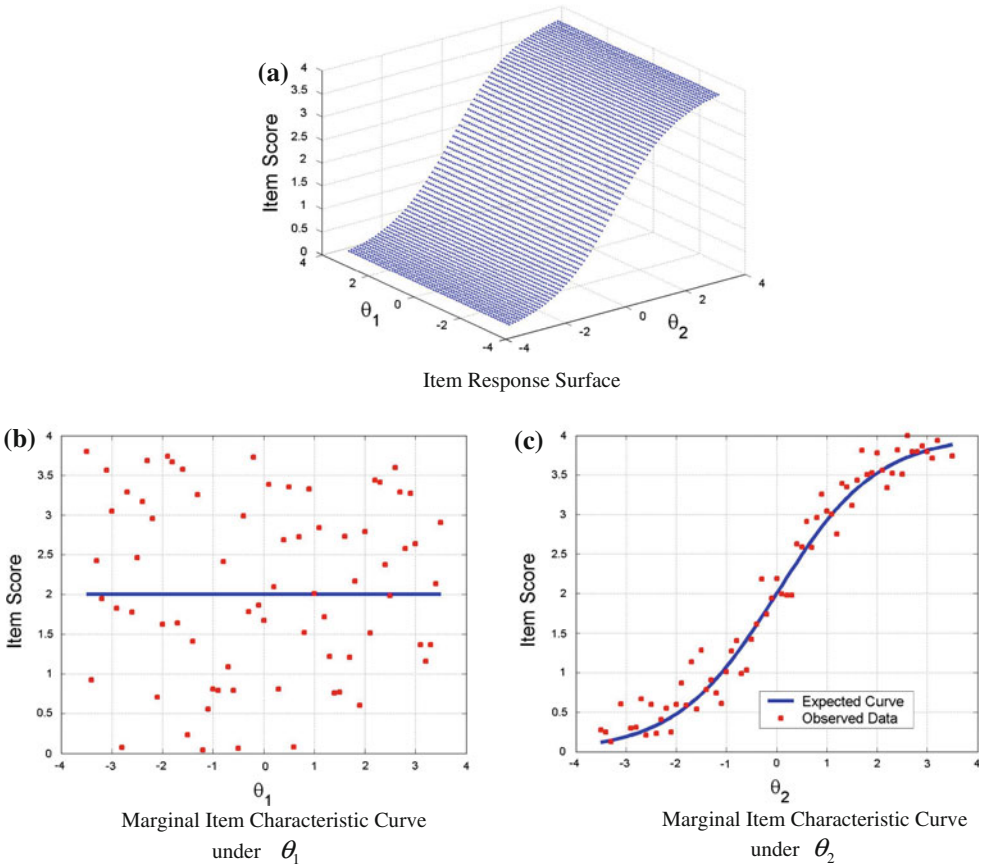

Fig. 1 Item response surface and item characteristic curves of an illustrative *M*-type polytomous item (z = 0, 1, 2, 3, and 4)

was used, and the related boundary characteristic curve is given by:

$$P^*_{ijx} = \frac{\exp[\alpha_{i1}\theta_{j1} + \alpha_{i2}\theta_{j2} + \delta_{xi}]}{1 + \exp[\alpha_{i1}\theta_{j1} + \alpha_{i2}\theta_{j2} + \delta_{xi}]}. \quad (3)$$

P^*_{ijx} denotes the boundary probability for examinee j to have a category score larger than x on item i. The *log-normal* (0, 0.5^2) distribution was used to obtain the generating discrimination parameters: α_{i1} for GRM-fitting item and α_{i2} for M-type misfit item. To ensure that each item measures only one construct, α_{i1} of the M-type misfit item measuring θ_2 was always zero. Likewise α_{i2} of the GRM-fitting item measuring θ_1 was always zero. For either the GRM-fitting or M-type misfit item, the procedures for generating β_1, β_2, β_3, and β_4 were the same process as those for the Type I error rate study. Then, the δ_{xi} values for M-type and GRM-fitting items are given by $-\alpha_{i2}\beta_{xi}$ and $-\alpha_{i1}\beta_{xi}$, respectively.

To generate a D-type misfit item, the boundary characteristic curves resembling the two-step Guttman functions as shown in Fig. 2 were used. The step functions had clear discrepancies from the GRM logistic curves. The item parameters for the boundary characteristic curves under the GRM were $\alpha = 2$, $\beta_1 = -2.1$, $\beta_2 = -0.7$, $\beta_1 = 0.7$, and $\beta_1 = 2.1$. In contrast to the generating procedure for the M-type misfit and GRM-fitting item data that used random selected values for generating item parameters, the generating procedure for the D-type misfit item data employed a fixed two-step Guttman functions in Fig. 2. (i.e., the distance between two adjacent Guttman curves was 1.4, and the middle points of the curves were fixed as −2.1, −0.7, 0.7, and 2.1, respectively).

For the power study, the main consideration was given to the effects of different types of misfit, sample size, and test length upon the empirical power of S-X^2. The number of item score categories was always 5. The standard normal distribution was used to generate abilities. Hence, the empirical power was examined under 18 simulated conditions (3 test lengths of 5, 10, and 20 × 3 sample sizes of 500, 1,000, and 2,000 × 2 types of misfit items). One hundred replications were generated for each condition. Each replication included a data set composed of I-1 items (I = 5, 10, or 20) simulated with the GRM functions and a single M- or D- type item. For each condition, the empirical power of the generalized S-X^2 index was computed using the number of times that the misfit items were correctly flagged.

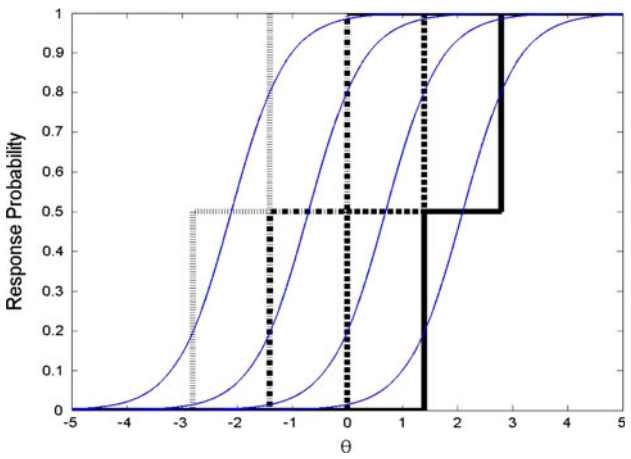

Fig. 2 Two-step Guttman functions of an illustrative D-type polytomous item (z = 0, 1, 2, 3, and 4) and the corresponding boundary characteristic curves of the GRM

Results

For the current study, the item parameters were estimated by the computer program MULTILOG 7.03 (Thissen 2003) using each simulated dataset. To calibrate the GRM parameters, the program defaults except N cycles = 3,000 were used. Most calibrations converged quickly, but sometimes over 1,000 EM cycles were required for the successful convergence.

Type I error rates

The proportions of items wrongly flagged for misfit are shown in Table 1 for each condition. The nominal alpha of 0.05 was used for the hypothesis tests in this paper. In the 5, 10, and 20 item test length conditions, the proportions were calculated based on the total of 500, 1,000, and 2,000 items, respectively (# items per dataset × 100 replications).

For the conditions with the generating distribution of N(0,1), the simulated Type I error rates tended to be closer to 0.05 than those of the conditions with U(−3,3) in most cases. In the N(0,1) conditions, the Type I error rates ranged from 0.030 to 0.078 except the condition with 20 item test, small sample size (500 examinees), and 5 item score categories. In a few cases of the conditions with U(−3,3), relatively large inflations of Type I error rates were observed, such as 0.120 and 0.225 when the sample size was 500.

The false rejection rates seemed to be influenced by the ratio of sample size to the number of test score groups. When there are 20 items with 5 categories in a test, the total number of groups is 73 (from $Z_i = 4$ to $F-Z_i = 76$). In these cases, the empirical Type I error rates were improved from 0.203 to 0.058 as the sample size increased from 500 to 2,000 in the N(0,1) conditions. The same improvement was found in the U(−3,3) conditions, where the simulated Type I error rates changed from 0.225 to 0.056. In other words, for the generalized S-X^2 to work properly, more data were required as the number of homogeneous test score groups increased.

To account for sampling error in obtaining the empirical Type I error rates under each simulation condition given a

Table 1 Type I error rates: proportions of indices with empirical *p*-values less than 0.05

Test length	Sample size	# Category	In conditions with $N(0,1)$	In conditions with $U(-3,3)$
5	500	3	0.046	0.068
		5	0.046	0.080*
	1,000	3	0.040	0.036
		5	0.056	0.064
	2,000	3	0.030*	0.054
		5	0.064*	0.058
10	500	3	0.054	0.073*
		5	0.060	0.120*
	1,000	3	0.044	0.040
		5	0.042	0.060
	2,000	3	0.046	0.042
		5	0.053	0.048
20	500	3	0.071*	0.099*
		5	0.203*	0.225*
	1,000	3	0.052	0.052
		5	0.078*	0.092*
	2,000	3	0.061*	0.062*
		5	0.058	0.056

*Indicates the empirical Type I error rate is out of 95% CI $= 0.05 \pm 1.96\sqrt{(0.05 \times 0.95)/100 \times I}$ where $I =$ test length

nominal alpha of 0.05, 95% confidence intervals were computed using CI = $0.05 \pm 1.96\sqrt{(0.05 \times 0.95)/(R \times I)}$ where R is the number of replication for each condition I is the test length (Stone and Hansen 2000; Zhang and Stone 2008). For the test length of 5, 10, and 20 and $R = 100$, the intervals were (0,031, 0.069) (0.036, 0.064), and (0.040, 0.060), respectively. Six conditions under the N(0,1) and seven conditions under the U(−3,3), appeared to have empirical Type I error rates falling out of the CIs. Because the extent of the Type I error rate inflation was usually small for those conditions, however, S-X^2 seemed to adequately control the Type I error rates in most cases.

Empirical power

Because each data set included only one misfit item in power study, the empirical power was calculated as the number of correctly flagged items divided by 100 in each condition. Table 2 summarizes the empirical power of the generalized S-X^2 in detecting *M*- or *D*-type misfit items. The false alarm rate (FAR) which indicates the proportions of the GRM-fitting items being wrongly flagged as misfit is reported in Table 2. FAR is a very similar concept to the empirical Type I error rate, but may be affected by true misfit items in the parameter estimation process.

As shown in Table 2, the empirical power for detecting the *M*-type misfit item ranged from 0.050 to 0.130. The FAR appeared to be between 0.038 and 0.079, which was similar to that of the Type I error rate study. Table 3 shows the average item parameter estimates and their standard deviations for the datasets including *M*-type misfit items. As discussed earlier in this paper, the estimated, the estimated discrimination parameters of *M*-type misfit items were close to zero. The average estimates were between 0.10 and 0.13 while the average discrimination parameter estimates of the GRM fitting items were between 1.09 and

Table 2 Empirical power rates: proportions of indices with empirical *p*-values less than 0.05

Misfit type	Test length	Sample size	Empirical power (FAR)
M	5	500	0.050 (0.043)
		1,000	0.120 (0.038)
		2,000	0.100 (0.048)
	10	500	0.120 (0.054)
		1,000	0.110 (0.053)
		2,000	0.120 (0.042)
	20	500	
		1,000	0.080 (0.079)
		2,000	0.130 (0.057)
D	5	500	0.110 (0.038)
		1,000	0.180 (0.048)
		2,000	0.410 (0.078)
	10	500	0.280 (0.059)
		1,000	0.320 (0.051)
		2,000	0.640 (0.072)
	20	500	
		1,000	0.830 (0.085)
		2,000	0.970 (0.057)

For conditions where the empirical Type I error rates were considerably inflated, the empirical power was not reported

1.28. Because an *M*-type item measured a very different construct from that measured by the other items in the same test, the threshold parameters were not able to be estimated accurately. This fact was reflected very well through the relatively much larger standard deviations of the *M*-type items' threshold parameters than those of the GRM-fitting items. Accordingly, it was expected that the *M*-type items should be easily detected as misfitting. However, the power of the generalized S-X^2 in finding the *M*-type misfit items was very small as shown in the results of this power study.

In detecting the *D*-type misfit items, as shown in Table 2, the power of the generalized S-X^2 appeared to increase as the sample size increased or as the test length became longer. For example, when the test length was 20, the empirical power of the item-fit statistics increased from 0.830 to 0.970 as the sample size increased from 1,000 to 2,000. Also when the sample size was 2,000, the empirical power estimates were 0.410, 0.650, and 0.970 for the test lengths of 5, 10, and 20, respectively.

Table 4 shows the average item parameter estimates and their standard deviations for the datasets including the *D*-type misfit items. In all conditions, the average threshold parameter estimates of the *D*-type misfit items were close to −2.1, −0.7, 0.7, and 2.1, which were the threshold parameters of the corresponding GRM curves in Fig. 2. The estimated discrimination parameters ($\hat{\alpha}_{1i}$) appeared a little larger (from 2.49 to 2.78) than the discrimination parameter (2.00) of the corresponding GRM curves in Fig. 2.

Discussion and conclusion

As more high-stake assessments adopt polytomous items analyzed by IRT, the need for polytomous IRT fit indices increases. As mentioned earlier, the success of IRT applications requires satisfactory fit between the model and the data. Also one of the most important features of IRT, parameter invariance, may not hold when the model is incorrect (Andersen 1973; Shepard et al. 1984). In answer to the call for an appropriate item fit index under polytomous IRT, the results of the current study provide a rationale for the application of the generalized S-X^2 as an item fit index under the GRM for researchers and practitioners.

The results of the Type I error rates study showed that the Type I error rates of the generalized S-X^2 adequately controlled around the nominal level, 0.05, in most conditions. In the conditions with the 20 items having 5 categories and small sample size (500 examinees), somewhat severe inflations of the Type I error rates were found. These results are consistent to those reported in Kang and Chen (2008). Also, as Kang and Chen mentioned, this can be explained by the sparseness problem in expected frequencies. The expected cell frequencies will be easily sparse when there is a small group of examinees given that there are many test score groups. Because more sparseness causes more collapsing, the *df* of the item-fit statistic would consequently be smaller. Then, it would be much easier to reject the null hypothesis of model-fit.

The power of the S-X^2 item-fit statistic was much lower when the source of misfit was *M*-type compared to *D*-type.

Table 3 Average item parameter estimates (SD) of the datasets including the *M*-type misfit items

Test length	Sample size	Item (#)	$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
5	500	*M*-type (100)	0.12 (0.04)	−7.47 (6.25)	−3.06 (5.22)	1.16 (5.17)	5.47 (5.81)
		GRM-fitting (400)	1.28 (2.45)	−0.98 (0.67)	−0.43 (0.64)	0.13 (0.63)	0.70 (0.63)
	1,000	*M*-type (100)	0.12 (0.03)	−7.19 (4.57)	−3.18 (4.53)	0.69 (4.98)	4.78 (5.91)
		GRM-fitting (400)	1.21 (1.16)	−1.05 (0.61)	−0.50 (0.61)	0.05 (0.60)	0.61 (0.61)
	2,000	*M*-type (100)	0.13 (0.03)	−6.36 (3.61)	−2.56 (3.56)	1.23 (4.58)	5.08 (6.16)
		GRM-fitting (400)	1.09 (0.53)	−0.98 (0.62)	−0.42 (0.62)	0.12 (0.63)	0.67 (0.65)
10	500	*M*-type (100)	0.12 (0.03)	−7.11 (4.69)	−2.89 (4.65)	1.57 (5.05)	5.81 (6.09)
		GRM-fitting (900)	1.16 (0.66)	−1.03 (0.62)	−0.48 (0.62)	0.09 (0.63)	0.65 (0.65)
	1,000	*M*-type (100)	0.12 (0.04)	−7.72 (4.54)	−3.44 (4.79)	0.58 (5.58)	4.72 (6.90)
		GRM-fitting (900)	1.15 (0.61)	−1.00 (0.59)	−0.44 (0.58)	0.11 (0.59)	0.66 (0.59)
	2,000	*M*-type (100)	0.13 (0.04)	−6.63 (3.48)	−2.91 (3.38)	0.89 (4.12)	4.83 (5.51)
		GRM-fitting (900)	1.14 (0.61)	−1.00 (0.61)	−0.44 (0.61)	0.12 (0.62)	0.67 (0.62)
20	500	*M*-type (100)	0.10 (0.03)	−8.08 (5.34)	−3.38 (4.83)	1.25 (5.24)	5.83 (6.21)
		GRM-fitting (1,900)	1.12 (0.60)	−1.05 (0.64)	−0.49 (0.63)	0.08 (0.63)	0.64 (0.64)
	1,000	*M*-type (100)	0.11 (0.02)	−7.80 (4.96)	−3.28 (5.04)	1.22 (5.62)	5.61 (6.86)
		GRM-fitting (1,900)	1.13 (0.61)	−1.00 (0.60)	−0.45 (0.60)	0.11 (0.61)	0.66 (0.63)
	2,000	*M*-type (100)	0.12 (0.04)	−6.60 (3.47)	−2.57 (4.33)	1.48 (6.11)	5.58 (8.04)
		GRM-fitting (1,900)	1.13 (0.62)	−0.99 (0.59)	−0.44 (0.59)	0.11 (0.60)	0.67 (0.61)

Table 4 Average item parameter estimates (SD) of the datasets including the *D*-type misfit items

Test length	Sample size	Item (#)	$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
5	500	*D*-type (100)	2.78	−2.12	−0.66	0.66	2.11
		GRM-fitting (400)	1.06 (0.60)	−1.10 (0.68)	−0.53 (0.63)	0.06 (0.64)	0.65 (0.69)
	1,000	*D*-type (100)	2.71	−2.11	−0.65	0.66	2.11
		GRM-fitting (400)	1.13 (0.70)	−1.04 (0.62)	−0.47 (0.60)	0.08 (0.60)	0.64 (0.61)
	2,000	*D*-type (100)	2.66	−2.11	−0.66	0.66	2.10
		GRM-fitting (400)	1.16 (0.64)	−1.05 (0.58)	−0.50 (0.58)	0.06 (0.58)	0.61 (0.59)
10	500	*D*-type (100)	2.57	−2.15	−0.67	0.67	2.14
		GRM-fitting (900)	1.12 (0.60)	−1.03 (0.65)	−0.46 (0.64)	0.10 (0.64)	0.67 (0.66)
	1,000	*D*-type (100)	2.61	−2.12	−0.66	0.67	2.12
		GRM-fitting (900)	1.12 (0.63)	−1.02 (0.59)	−0.47 (0.59)	0.09 (0.60)	0.65 (0.61)
	2,000	*D*-type (100)	2.60	−2.11	−0.65	0.66	2.12
		GRM-fitting (900)	1.12 (0.62)	−1.02 (0.58)	−0.46 (0.59)	0.09 (0.59)	0.64 (0.60)
20	500	*D*-type (100)	2.49	−2.18	−0.69	0.69	2.17
		GRM-fitting (1,900)	1.10 (0.59)	−1.02 (0.64)	−0.45 (0.63)	0.13 (0.64)	0.70 (0.65)
	1,000	*D*-type (100)	2.58	−2.11	−0.65	0.67	2.13
		GRM-fitting (1,900)	1.12 (0.60)	−1.00 (0.61)	−0.44 (0.61)	0.12 (0.62)	0.67 (0.63)
	2,000	*D*-type (100)	2.58	−2.10	−0.66	0.68	2.11
		GRM-fitting (1,900)	1.14 (0.61)	−1.00 (0.58)	−0.45 (0.58)	0.10 (0.59)	0.65 (0.59)

The small SD values of *D*-type items ranging between 0.04 and 0.20 due to the fixed data generating functions were not reported

Even though the *M*-type items appeared to cause the poor threshold parameter estimation, S-X^2 seemed insensitive to detect this type of misfit even with a large sample size of 2,000. Because previous studies (e.g., Orlando and Thissen 2003; Zhang and Stone 2008) reporting satisfactory empirical powers of S-X^2 statistics did not consider *M*-type but only *D*-type misfit items, the results here provide noteworthy information for future studies.

For the *M*-type misfit items, it was true that the empirical power was too low to suggest the S-X^2 statistic as a tool for evaluating item fit under all the conditions considered in this study. Also for the *D*-type misfit items, it appeared that appropriate power could be only obtained for conditions with a 20-item test and a sample size of at least 1,000. Because it is noted that the statistical power is a function of sample size, it can be expected that more power could be observed for larger sample size. Under this consideration, a further study was conducted to investigate appropriate sample sizes that would produce satisfactory power (e.g., 0.7 or higher). This additional power study employed much larger sample sizes of 5,000, 10,000, and 20,000 for detecting the *M*- or *D*-type misfit items under the conditions with a test length of 5, 10, or 20 items each having 5 categories. The number of these additional conditions was 18 (= 3 test lengths × 3 new sample sizes × 2 types of misfit items). Each data set included a single misfit item, and 100 replicated data sets for each new condition were generated and calibrated in the same process as described earlier. The results are summarized in Table 5.

As shown in Table 5, regardless of test length, it appeared that at least 20,000 examinees were required to obtain acceptable power in detecting misfit items due to multidimensionality. However, for a short test with five items, inflated FARs, 0.078 and 0.115, were found for the conditions with 10,000 and 20,000 examinees, respectively. For the *D*-type misfit items, the sample sizes of 5,000 or more appeared to be enough for producing satisfactory power. But, similar to the cases of M-type items, for a short test with five items, inflated FARs, 0.130 and 0.185, were found for the conditions with 10,000 and 20,000 examinees, respectively. These results indicate that the use of very large sample sizes should be considered more cautiously when the test length is short (five items) regardless of the source of item misfit.

In conclusion, under the GRM, the generalized S-X^2 adequately controlled the Type I error rates in most conditions and was able to yield satisfactory power in detecting the *M*- or *D*-type misfit items with large sample sizes. Therefore, the generalized S-X^2 item-fit statistics appeared to be a promising index for investigating item-fit in educational and psychological assessments having polytomous items.

To gain a better understanding of this promising item-fit index, however, additional studies need to be conducted. First, because this study examined only one kind of misfit item under each test data set, additional misfit items in a test need to be further considered to better understand the performance of the generalized S-X^2 in detecting misfit

Table 5 Empirical power rates of an additional power study with very large sample sizes: proportions of indices with empirical *p*-values less than 0.05

Misfit type	Test length	Sample size	Empirical power (FAR)
M	5	5,000	0.290 (0.048)
		10,000	0.470 (0.078)
		20,000	0.730 (0.115)
	10	5,000	0.170 (0.043)
		10,000	0.490 (0.056)
		20,000	0.740 (0.059)
	20	5,000	0.280 (0.053)
		10,000	0.460 (0.050)
		20,000	0.740 (0.046)
D	5	5,000	0.790 (0.060)
		10,000	0.950 (0.130)
		20,000	0.990 (0.185)
	10	5,000	1.000 (0.061)
		10,000	1.000 (0.064)
		20,000	1.000 (0.046)
	20	5,000	1.000 (0.054)
		10,000	1.000 (0.056)
		20,000	1.000 (0.055)

items. Also more potential sources of misfittng items (e.g., different levels of correlation between θ_1 and θ_2, differential item functioning, poorly estimated item parameters, etc.) need to be further studied. Second, different from the study of Kang and Chen (2008) where all the items on a test were misfit in conducting the power study, the current power study included only a single misfit item into each test. To make this approach more informative and meaningful, however, more different percentages or number of misfit items on a test needs to be further considered. Third, the performance of the generalized S-X^2 needs to be investigated under conditions where the ability distribution is other than normal and uniform. For example, under skewed distributions, it would be expected that S-X^2 performs poorly due to low cell frequencies in some parts of the distribution. Fourth, the Guttman step functions were used in this study to simulate the *D*-type misfit items. If empirical item response functions which are not consistent with the underlying model could be considered, it would make this type of study deal with more realistic misfit items. Finally, the test lengths considered in this study were 5, 10, and 20 items. Because it is not uncommon that a psychological scale includes more than 20 polytomous items, it will be also interesting to extend the current study into considering longer test lengths.

Acknowledgments This work was supported by the Sungshin Women's University Research Grant of 2010.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123–140.

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

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

Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X^2 item-fit index for polytomous IRT models. *Journal of Educational Measurement, 45*, 391–406.

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*, 412–432.

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*, 156–166.

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

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.

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.

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.

Thissen, D. (2003). *MULTILOG 7: Multiple categorical item analysis and test scoring using item response theory [computer program]*. Chicago: Scientific Software Corporation.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567–577.

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.

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

Zhang, B., & Stone, C. A. (2008). Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement, 68*, 181–196.