Research Report

The Correlation Between Item Parameters and Item Fit Statistics

Sandip Sinharay

Ying Lu

Research & Development

August 2007
RR-07-36

The Correlation Between Item Parameters and Item Fit Statistics

Sandip Sinharay and Ying Lu

ETS, Princeton, NJ

August 2007

Abstract

Dodeen (2004) studied the correlation between the item parameters of the three-parameter logistic model and two item fit statistics, and found some linear relationships (e.g., a positive correlation between item discrimination parameters and item fit statistics) that have the potential for influencing the work of practitioners who employ item response theory. This paper examines the same type of linear relationships as studied in Dodeen. However, this paper adds to the literature by employing item fit statistics not considered in Dodeen, which have been recently suggested and whose Type I error rates have been demonstrated to be generally close to the nominal level. Detailed simulations show that if one uses certain of the recently suggested item fit statistics, there is no need to worry about any linear relationships between the item parameters and item fit statistics.

Key words: ANOVA, χ^2 test, model fit, residual, Type I error

Acknowledgments

The authors thank Shelby Haberman, Dan Eignor, Frank Rijmen, and Bernard Veldkamp for useful advice, and Amanda McBride and Kim Fryer for help with proofreading.

Model checking remains a major hurdle to the effective implementation of item response theory (IRT; Hambleton & Han, 2004). Recent works like Stone and Zhang (2003), Orlando and Thissen (2003), Hambleton and Han (2004), and Sinharay (2005) notwithstanding, there is substantial scope of further research needed on the topic. Item fit is a major area of interest in model checking. Though researchers have suggested several different item fit statistics (e.g., Bock, 1972; Orlando & Thissen, 2000; Sinharay, 2006; Stone, 2000; Stone & Zhang, 2003; Glas & Suarez-Falcon, 2003; Yen, 1981), there is a lack of sufficient knowledge regarding factors that usually cause item misfit. For example, more appropriate assessments have resulted because the substantial existing knowledge of factors affecting differential item functioning, or DIF, (see, for example, Schmitt, Holland, & Dorans, 1993, and the references therein) often help test developers to control the number of items with DIF. Unfortunately, there is a general lack of such knowledge regarding the factors affecting item misfit.

In an attempt to explore such factors, Dodeen (2004), in the context of the three-parameter logistic (3PL) model, studied the linear relationships between item parameters and two item fit statistics: (a) the G^2-like item fit statistic χ^2_G (Mislevy & Bock, 1990), and (b) the standardized residual (Hambleton, Swaminathan, & Rogers, 1991). The paper reported substantial linear relationships (e.g., a positive correlation between the discrimination parameters and the item fit statistics) and also a positive correlation between the guessing parameters and the item fit statistics. These findings have the potential to influence construction of assessments that employ IRT models. For example, the positive correlation between the discrimination parameters and item fit statistics that Dodeen found may create a dilemma regarding the use of highly discriminating items in tests.

Several item fit statistics have been suggested recently, by researchers such as Glas and Suarez-Falcon (2003), Orlando and Thissen (2000), Sinharay (2006), Stone (2000), and Stone and Zhang (2003). There is a need to study the same relationships as studied by Dodeen (2004), but with these recently developed item fit statistics; if the relationships hold for these newer statistics as well, there will be sufficient reason to be careful about test construction.

Hence, this paper examines the same relationships studied by Dodeen (2004) using several simulated data sets and a real data set employing several newer item fit statistics: The $S-\chi^2$ and $S-G^2$ statistics of Orlando and Thissen (2000) and the χ^{2^*} and G^{2^*} statistics of Stone (2000).

These four statistics, unlike those used by Dodeen, have been demonstrated to have Type I error rates generally close to the nominal level under a wide variety of conditions. The first two of these use examinee groups defined using the raw score scale while the latter two use examinee groups defined using the ability parameter scale. This paper performs simulations under the same conditions as in Dodeen, and also under more conditions.

The next section describes the study by Dodeen (2004). The Simulations section covers simulations like in Dodeen. The Further Simulations section shows results from simulations under more conditions than considered in Dodeen. The Real Data section discusses results from a real data example. The Closer Look section examines the reasons behind the differences between Dodeen's results and those in this paper. The last section provides discussion and recommendations.

Brief Description of the Study of Dodeen

Dodeen (2004) studied the linear relationships between item parameters and item fit statistics for data generated from and analyzed using the 3PL model. The author employed two item fit statistics. The first is χ^2_G (Mislevy & Bock, 1990)[1] given by:

$$\chi^2_G = 2\sum_{j=1}^{n} N_j \left[O_j \log\left(\frac{O_j}{E_j}\right) + (1-O_j)\log\left(\frac{1-O_j}{1-E_j}\right)\right], \tag{1}$$

where the ability (θ) scale is divided into n groups; O_j and E_j are, respectively, the observed and expected proportions of correct responses to the item in ability group j; and N_j is the number of examinees in group j. The second item fit statistic used by Dodeen is the standardized residual (SR; Hambleton, Swaminathan, & Rogers, 1991) given by:

$$z_j = \frac{(O_j - E_j)}{\sqrt{E_j(1-E_j)/N_j}}, j = 1, 2, \ldots n\cdot \tag{2}$$

Under each of nine test conditions, Dodeen (2004) simulated and analyzed 100 data sets, each with 1,000 examinees and 50 multiple-choice items, employing the 3PL model. Examinee ability parameters were generated from a $\mathcal{N}(0,1)$ distribution. The item parameters under the different test conditions were drawn from a normal distribution with means and standard deviations (SD) as shown in Table 1. Note that the first three test conditions differ only in mean discrimination, the next three only in mean difficulty and the last three only in mean discrimination. BILOG 3.11 (Mislevy & Bock, 1990) was used for fitting the 3PL model to the simulated data sets and for computing the item fit statistics.

Table 1
Generating Item Parameter Distributions

Test	Discrimination(a)		Difficulty(b)		Guessing(c)	
condition	Mean	SD	Mean	SD	Mean	SD
1	0.5	0.5	0.0	1.0	0.1	0.1
2	1.0	0.5	0.0	1.0	0.1	0.1
3	1.5	0.5	0.0	1.0	0.1	0.1
4	1.0	0.5	-1.0	1.0	0.1	0.1
5	1.0	0.5	0.0	1.0	0.1	0.1
6	1.0	0.5	1.0	1.0	0.1	0.1
7	1.0	0.5	0.0	1.0	0.1	0.1
8	1.0	0.5	0.0	1.0	0.25	0.1
9	1.0	0.5	0.0	1.0	0.5	0.1

Dodeen (2004) studied the linear relationships between item parameters and item fit statistics in several ways. The paper reported average item fit statistics, and the correlations between the item parameters and the average item fit statistics (averaged over 100 replications) under each of the nine test conditions in Table 1. Further, an analysis of variance (ANOVA) followed by pairwise comparisons with the average item fit statistics as the dependent variable and the test condition as the independent variable was conducted on the first three test conditions (to study the effect of the discrimination parameters on the item fit statistics), on the second three test conditions (to study the effect of the difficulty parameters on the item fit statistics), and on the last three test conditions (to study the effect of the guessing parameters on the item fit statistics).

Dodeen found for both of the two item fit statistics, χ^2_G and z_j, that the average, proportion significant, and the correlation with item parameters increased with an increase in the average discrimination parameters, and also with an increase in the average guessing parameters. No such phenomenon was observed for the difficulty parameters. From these results, Dodeen concluded that there is a positive correlation between the item discrimination parameters and both item fit statistics, and also between the item guessing parameters and both item fit statistics.

The findings of Dodeen (2004) may have serious consequences for constructing assessments that employ IRT models. Items with high discrimination parameters have high values of information and are usually preferred over other types of items, especially in computer-adaptive tests (CAT; e.g., Leung, Chang, & Hau, 2002). So the positive correlation between the discrimination parameters and item fit statistics that Dodeen found may create a dilemma regarding the use of highly discriminating items. Some practitioners might subject the highly discriminating items (assuming these to be more prone to item misfit) to more review than is necessary, or remove such items from the item pool, which would result in increased cost. On the other hand, some other practitioners might always retain highly discriminating items in the operational item pool and ignore observed misfit of such items (using Dodeen's finding to conclude that such items, even if from the correct model, have an increased tendency of showing misfit); this is not a good strategy because the item pool may have truly misfitting items that are highly discriminating, and retaining those items in the item pool would lead to tests with less than desirable properties.

Simulations Under the Test Conditions Considered by Dodeen (2004)

We first performed simulations under the nine test conditions considered by Dodeen (2004), but with five item fit statistics, one of which was employed by Dodeen.

The Item Fit Statistics Considered

The χ^2_G statistic was included in the simulations, as in Dodeen (2004), as well as the $S-\chi^2$ and $S-G^2$ statistics suggested by Orlando and Thissen (2000). For computing $S-\chi^2$ and $S-G^2$, the examinees were divided into G groups based on their raw scores. The $S-\chi^2$ statistic is given by

$$S-\chi^2 = \sum_{j=1}^{n} \frac{N_j(O_j - E_j)^2}{E_j(1-E_j)}, \tag{3}$$

and the $S-G^2$ statistic is given by

$$S-G^2 = 2\sum_{j=1}^{n} N_j \left[O_j \log\left(\frac{O_j}{E_j}\right) + (1-O_j)\log\left(\frac{1-O_j}{1-E_j}\right)\right], \tag{4}$$

where, O_j and E_j are the observed and expected proportions of correct responses, respectively, to the item in raw score group j, and N_j is the number of examinees in raw score group j. Glas and Suarez-Falcon (2003), Orlando and Thissen (2000), Sinharay (2006), and Stone and Zhang (2003) used detailed simulations to show that when the 3PL model is fit to the data, $S-\chi^2$ and $S-G^2$

are distributed approximately as a χ^2 random variable with $n-4$ degrees of freedom. The two statistics have slightly inflated Type I error rates for short tests (Glas & Suarez-Falcon, 2003; Sinharay, 2006).

Two additional statistics considered in this paper, suggested by Stone (2000), use a predetermined number of examinee groups defined on the scale of the examinee proficiency parameter θ. One computes the posterior probability for each examinee of belonging to each group. Then, for each item and each examinee group, one computes the *observed* number of examinees (often called *pseudo-counts* because these numbers are not truly observed) in each group who answered the item correctly/incorrectly, and the corresponding *expected* numbers. Then, one computes a χ^2-type and a G^2-type statistic, comparing the observed and expected proportions using formulae similar to Equations 1 and 3, respectively. Research has shown that the fit statistic is a scaled χ^2 random variable (Stone, 2000). To estimate the scaling factor and the effective degree of freedom, a resampling-based procedure is used that rescales the χ^2-type and G^2-type statistics to conform to a known χ^2 distribution for hypothesis testing. These rescaled statistics are henceforth denoted as χ^{2^*} and G^{2^*}, respectively. Several studies (Stone, 2000; Stone & Hansen, 2000; Stone & Zhang, 2003) found these statistics to have Type I error rates close to the nominal level and adequate power. Lu and Lin (2005) found occasionally high Type I error rates for these statistics.

While the two item fit statistics suggested by Orlando and Thissen (2000) are computed using ability groups based on the raw scores of examinees, the fit statistics of Stone (2000) are computed using ability groups on the proficiency scale. These are the two major ways of forming ability groups, and hence the item fit statistics chosen for use in this paper are representatives of the range of recently suggested item fit statistics. Also, these statistics are arguably the most popular ones in the psychometrics literature and have been shown to perform satisfactorily for a wide variety of conditions.

Study Design

We simulated and analyzed 100 data sets, each with 1,000 examinees and 50 multiple choice items, under each of the nine test conditions shown in Table 1, much in the same way as in Dodeen (2004). Note that Test Conditions 5 and 7 are the same as Test Condition 2. As in

Dodeen, examinee abilities were always generated from a $\mathcal{N}(0, 1)$ distribution, and the item parameters for the different test conditions were drawn from distributions with means and SDs as given in Table 1. However, Dodeen used a normal distribution for generating item parameters, which could lead to negative values of discrimination and guessing parameters in some cases. Dodeen did not discuss how the negative values were handled. To prevent the occurrence of negative values, we used a log-normal distribution for generating discrimination parameters and a beta distribution for generating guessing parameters. The parameters of the log-normal and beta distributions were chosen to make the mean and SD of the generating distributions the same as those in Table 1. The values of the generating item parameters remained the same for the 100 data sets generated under any test condition (another version of the simulations allowed the generating item parameters to vary over the 100 data sets, but the conclusions were the same—so those results are not reported).

As in Dodeen (2004), the BILOG 3.11 software (Mislevy & Bock, 1990) was used for fitting the 3PL model to the generated data sets and for computing the values of the χ^2_G statistic. The statistics $S - \chi^2$ and $S - G^2$ were computed using the GOODFIT (Orlando, 1997) program. The statistics χ^{2*} and G^{2*} were computed using the IRTFIT_RESAMPLE program (Stone, 2004).

As in Dodeen (2004), the average item fit statistics, the proportion of item fit statistics that are significant at 1% level and the correlations between the generating item parameters and the average item fit statistics (averaged over the 100 replications under any test condition) were computed under each of the nine test conditions. As in Dodeen, to determine the effect of each parameter level on the average item fit statistics, an analysis of variance (ANOVA) followed by a pairwise comparison was performed with the average item fit statistics as the response variable for Test Conditions 1-3 (to study the effect of discrimination parameters on item fit statistics), then for Test Conditions 4-6 (to study the effect of difficulty parameters on item fit statistics), and finally for Test Conditions 7-9 (to study the effect of guessing parameters on item fit statistics).

Results

Table 2 summarizes the results of the simulations for the $S - \chi^2$, χ^{2*}, and χ^2_G statistics. Because the G^2-type statistics produced very similar results as the corresponding χ^2-type statistics, values for the $S - G^2$ and G^{2*} statistics are not shown. In the table, the correlations

Table 2

Average Values, Proportion of Misfits (at 1% Level), and Correlations Between Item Parameters and Item Fit Statistics for $S-\chi^2$, χ^{2*} and χ^2_G for the Nine Test Conditions

	Test Condition								
	1	2	3	4	5	6	7	8	9
$S-\chi^2$: Av	27.8	32.9	31.3	30.3	32.9	32.5	32.9	29.6	23.2
$S-\chi^2$: Prop	0.01	0.01	0.03	0.01	0.01	0.01	0.01	0.01	0.01
$S-\chi^2$: Cor	-0.32	-0.40*	-0.32	0.08	0.15	0.30	0.09	0.08	0.00
χ^{2*} : Av	3.0	4.7	5.1	4.0	4.7	4.3	4.7	3.0	3.3
χ^{2*} : Prop	0.01	0.03	0.04	0.01	0.03	0.04	0.03	0.01	0.02
χ^{2*} : Cor	-0.17	-0.11	-0.14	0.13	0.04	0.13	-0.16	0.06	-0.03
χ^2_G: Av	9.5	12.1	16.9	8.1	12.1	11.8	12.1	8.2	8.7
χ^2_G: Prop	0.03	0.10	0.34	0.03	0.10	0.09	0.10	0.03	0.02
χ^2_G: Cor	0.32	0.25	0.15	0.49*	0.13	-0.31	-0.22	0.03	0.01

Note. "Av" denotes average statistic, "Prop" denotes proportion significant at 1% level, and "Cor" denotes correlation.

reported for Test Conditions 1-3 are between the average discrimination parameters and the average item fit statistics, the correlations for Test Conditions 4-6 are between the average difficulty parameters and the average item fit statistics, and the correlations for Test Conditions 7-9 are between the average guessing parameter and the average item fit statistics. The correlation coefficients that are significant at the 1% level (using the result that $\sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \sim t_{n-2}$ for a bivariate normal distribution with population correlation coefficient 0; see, e.g., Rohatgi, 1976) are marked with an asterisk in the table.

Results for χ^2_G. Relationships between values of the χ^2_G statistic and those of the slope parameters are somewhat similar to those observed in Dodeen (2004). The average and proportion significant for χ^2_G increases with an increase in the average discrimination parameter (i.e., over Test Conditions 1-3). However, unlike in Dodeen (2004), the correlation decreased with an increase in the average discrimination parameter. Relationships between the difficulty parameters and the values of the χ^2_G statistics were somewhat different from those in Dodeen, but no consistent pattern was found in the average, proportion significant, and the correlation for χ^2_G over Test Conditions 4-6. Unlike Dodeen's results, our results do not show any linear relationships between the guessing parameters and the χ^2_G statistic: The statistic remains unaffected by an increase in the average guessing parameter. We wonder whether a reason for the differences between the results here and those in Dodeen is that Dodeen used the normal distribution for generating the discrimination and guessing parameters, which might have led to a substantial number of negative values of these parameters.

Results for $S - \chi^2$ and χ^{2^*}. The proportion significant for the $S - \chi^2$ and χ^{2^*} are low and close to the nominal level, for all the test conditions. For the $S - \chi^2$ statistic, the averages, the proportions significant, and the correlations do not show any pattern like those in Dodeen (2004). Note the substantial negative correlations between average $S - \chi^2$ and the slope parameter for Test Conditions 1-3. The negative correlations should not cause much worry because, even if they indicate any causal relationship, test developers generally try to include items with high discrimination parameters anyway, which automatically keeps the values of the $S - \chi^2$ statistic in control. The average χ^{2^*} statistic increases somewhat over Test Conditions 1-3 (suggesting that it increases somewhat with an increase in the slope parameter), but the corresponding proportion significant and the corresponding correlation do not show any consistent pattern over the test conditions. ANOVA tests of the same kind performed in Dodeen (2004) do not reveal any linear relationships between item parameters and any of these two item fit statistics. For the $S - \chi^2$ statistic, the three ANOVA tests resulted in nonsignificant values of 0.84 (p-value = 0.43), 0.29 (p-value = 0.74), and 0.05 (p-value = 0.95) of the F-statistic with degrees of freedom (df) of 2 and 147 (as there were three test conditions and 150 total items for the three test conditions combined for each ANOVA test). For the χ^{2^*} statistic, the corresponding values of the F-statistic were 2.17 (p-value = 0.12), 2.42 (p-value = 0.09), and 0.96 (p-value = 0.37), respectively. So, based on these simulations, the $S - \chi^2$ and χ^{2^*} statistics do not show any pattern that will cause any worry to IRT practitioners.

Further Simulations

The above simulations, as in Dodeen (2004), considered only one sample size (1,000) and one test length (50). However, as sample size and test length differ, item fit statistics often exhibit different and occasionally poor Type I error rates. For example, Orlando and Thissen (2000) found the Type I error rate of the $S - G^2$ statistic to vary between 0.09 and 0.13 at the 5% level for Test

Lengths 10, 40, and 80, while Glas and Suarez-Falcon (2003) found the Type I error rate of the $S - G^2$ statistic to be greater than or equal to 0.07 at the 5% level for 10-item tests, irrespective of the sample size. Therefore, this section examines the linear relationships between item parameters and item fit statistics for several sample sizes and test lengths. Three test lengths, 10, 30, and 50 (representing short, medium, and long tests, respectively), were considered, as well as three sample sizes, 500, 1,000, and 2,000 (representing small, medium, and large samples, respectively). Note that simulations for 1,000 examinees and 50 items were performed earlier. For each of the nine combinations of sample size and test length, 100 data sets each were generated under each of the first three test conditions described earlier. The first three test conditions are adequate for studying the linear relationship between the item discrimination parameters and the item fit statistics, which is of prime concern in this paper.

As in Dodeen (2004), examinee abilities were generated from a $\mathcal{N}(0,1)$ distribution. The item parameters were randomly drawn from distributions with means and SDs as given in Table 1 in the same manner as in the earlier simulations. The values of the generating item parameters remained the same for the 100 data sets generated under any simulation condition.

As in Dodeen (2004), the BILOG 3.11 software (Mislevy & Bock, 1990) was used for fitting the 3PL model to the generated data sets and for computing the χ^2_G statistic. The $S - \chi^2$, $S - G^2$, χ^{2*}, and G^{2*} statistics were computed using the GOODFIT (Orlando, 1997) and IRTFIT_RESAMPLE (Stone, 2004) software, respectively.

Table 3 shows the average and percent significant (at the 1% level) for the χ^2_G, $S - \chi^2$, and χ^{2*} statistics for the first three test conditions for each of the nine combinations of test length and sample size.

To systematically study if the test conditions affect the item fit statistics, we performed ANOVAs with each of the six quantities (average and percent significant for the three item fit indices) reported in the last six columns of Table 3 as the dependent variable, and the test length, sample size, and test condition as the three independent variables. Because the values of the average χ^2_Gs are often high for Test Length 10, we performed the ANOVAs on the logarithm of the average item fit statistics.

Table 3 and the ANOVA results indicate the following on the Type I error rates (or the percent significant) for the item fit statistics:

Table 3

Average Values and Percent of Misfits Between Item Parameters and Item Fit Statistics for χ^2_G, $S-\chi^2$, and χ^{2*} for Several Simulation Conditions

Test	Sample	Test	χ^2_G		$S-\chi^2$		χ^{2*}	
length	size	condition	Av	%	Av	%	Av	%
10	500	1	21.6	49	6.8	5	0.8	8
		2	50.2	75	5.5	1	1.3	5
		3	59.7	68	5.8	2	1.7	5
	1,000	1	38.1	67	7.4	4	1.1	5
		2	63.8	75	6.5	0	0.9	2
		3	187.2	83	6.4	0	1.4	8
	2,000	1	37.7	72	9.1	1	4.1	53
		2	57.9	72	8.8	4	3.6	36
		3	99.7	74	9.3	3	5.2	37
30	500	1	9.2	5	16.9	1	2.0	2
		2	8.0	3	19.6	1	2.7	1
		3	9.1	7	17.9	0	3.3	3
	1,000	1	11.7	12	18.4	1	2.0	4
		2	10.8	9	20.7	1	3.0	2
		3	12.8	16	20.1	2	3.9	3
	2,000	1	11.1	10	21.0	2	5.5	30
		2	11.2	10	23.7	2	6.9	19
		3	13.0	20	24.4	3	8.0	23
50	500	1	8.6	4	25.4	1	3.1	3
		2	8.3	2	29.0	2	4.1	2
		3	9.1	8	26.4	1	4.6	3
	1,000	1	9.5	3	27.8	1	3.0	1
		2	12.1	10	32.9	1	4.7	3
		3	16.9	34	31.3	3	5.1	4
	2,000	1	9.6	4	32.8	3	6.2	18
		2	11.0	7	38.3	3	8.0	14
		3	17.0	32	37.7	3	9.8	20

Note. "Av" denotes average statistic, and "%" denotes the percentage of item fit statistics significant at 1% level.

- The percent significant for the χ^2_G statistic is generally much higher than the nominal level of 1%, which suggests that the statistic should not be used to evaluate item fit.
- The percent significant for χ^{2*} is close to the nominal 1% level for sample sizes 500 and 1,000, but very high for sample size 2,000. This finding suggests the need for further research

regarding χ^{2*}. Research by Stone and colleagues (e.g., Stone, 2000; Stone & Hansen, 2000; Stone & Zhang, 2003) demonstrated respectable Type I error rates of χ^{2*} under a wide variety of conditions, but these papers did not fit the 3PL model. On the other hand, Lu and Lin (2005) found occasionally high Type I error rates of the χ^{2*} and G^{2*} statistics when the 3PL model was fitted to data generated from the 3PL model.

- The Type I error rate for the $S-\chi^2$ statistic is always close to the nominal level of 1% (and almost always lowest among the three statistics considered in Table 3) for these simulations.

Table 3 and the ANOVA results indicate the following on the effect of the slope parameters on average item fit statistics and the percent significant for average fit statistics:

- The χ^2_G statistic is affected by test condition. Higher average slope parameters generally result in higher average and higher percent significant for χ^2_G. The main effect of test condition is statistically significant in the ANOVA for either of average χ^2_G or percent significant for χ^2_G as the dependent variable. This effect is the most prominent for Test Length 10 (there is a sharp rise in average χ^2_G for Test Conditions 2 and 3 for Sample Sizes 1,000 and 2,000) followed by Test Length 50.

- The average value of χ^{2*} is affected by the test condition. The main effect of test condition is statistically significant when average χ^{2*} is the dependent variable. In Table 3, the average χ^{2*} often increases with an increase in the average slope parameter. The main effect of test condition is not statistically significant when percent significant for χ^{2*} is the dependent variable.

- The statistic $S-\chi^2$ is not affected by test conditions. The main effect of test condition is not statistically significant for either the average $S-\chi^2$ or the percent significant for $S-\chi^2$ as the dependent variable.

Thus, our simulations support the result of Dodeen (2004) that higher values of the average slope parameter result in higher values of the χ^2_G statistic. The same effect is also noticed to some extent for the χ^{2*} statistic. However, no such effect is observed for the $S-\chi^2$ statistic. Besides, the $S-\chi^2$ statistic has Type I error rates close to the nominal level in our simulations. Hence, the simulations demonstrate the superiority of the $S-\chi^2$ statistic over the other statistics considered.

A Real Data Example

Next, we examine the linear relationships between item parameter estimates and item fit statistics for a real item response data set. The data set, from a basic skills test considered in Sinharay (2005), has 8,686 examinees and 45 multiple-choice items. Figure 1 shows the values of the $S-\chi^2$ (top row), χ^{2^*} (middle row), and χ^2_G (bottom row) statistics versus the item parameter estimates obtained using BILOG 3.11 (Mislevy & Bock, 1990) from the data set. Each plot shows the corresponding correlation coefficients (denoted as *Corr*) between the item parameter estimates and the item fit statistics.

Results for χ^2_G. At the 1% level, there are 21 significant values of the χ^2_G statistic, which clearly reflects its inflated Type I error rate. While there is a positive correlation between the estimated discrimination parameters and the χ^2_G statistic, there is a negative correlation between the estimated difficulty parameters and the values of the χ^2_G statistic, and also between the estimated guessing parameters and the values of the χ^2_G statistic. The last two of these three correlations are statistically significant at the 1% level. Negative correlations for the χ^2_G statistic were not found in the simulations in Dodeen (2004) and were seen rarely in this study. However, the simulations were for the situation when the true model is the 3PL model, while the true model is unknown for these real data.

Results for $S-\chi^2$ *and* χ^{2^*}. At the 1% level, there are two significant values for the $S-\chi^2$ statistic, and no significant value for the χ^{2^*} statistic. There is a negative nonsignificant correlation between the estimated discrimination parameters and the item fit statistics for both $S-\chi^2$ and χ^{2^*}. The correlations are of opposite signs, and both nonsignificant, between the estimated guessing parameters and the $S-\chi^2$ and χ^{2^*} statistics. The same is true for the estimated difficulty parameters. Also, a multiple regression analysis of the values of the $S-\chi^2$ statistic on the estimated discrimination, difficulty, and guessing parameters resulted in a squared multiple correlation coefficient of only 0.07 and an F-statistic (with df of 3 and 41) with p-value = 0.37; the corresponding values for the χ^{2^*} statistic are 0.08 and 0.29. Thus, there are no obvious linear relationships between the item parameter estimates and either of these two statistics.

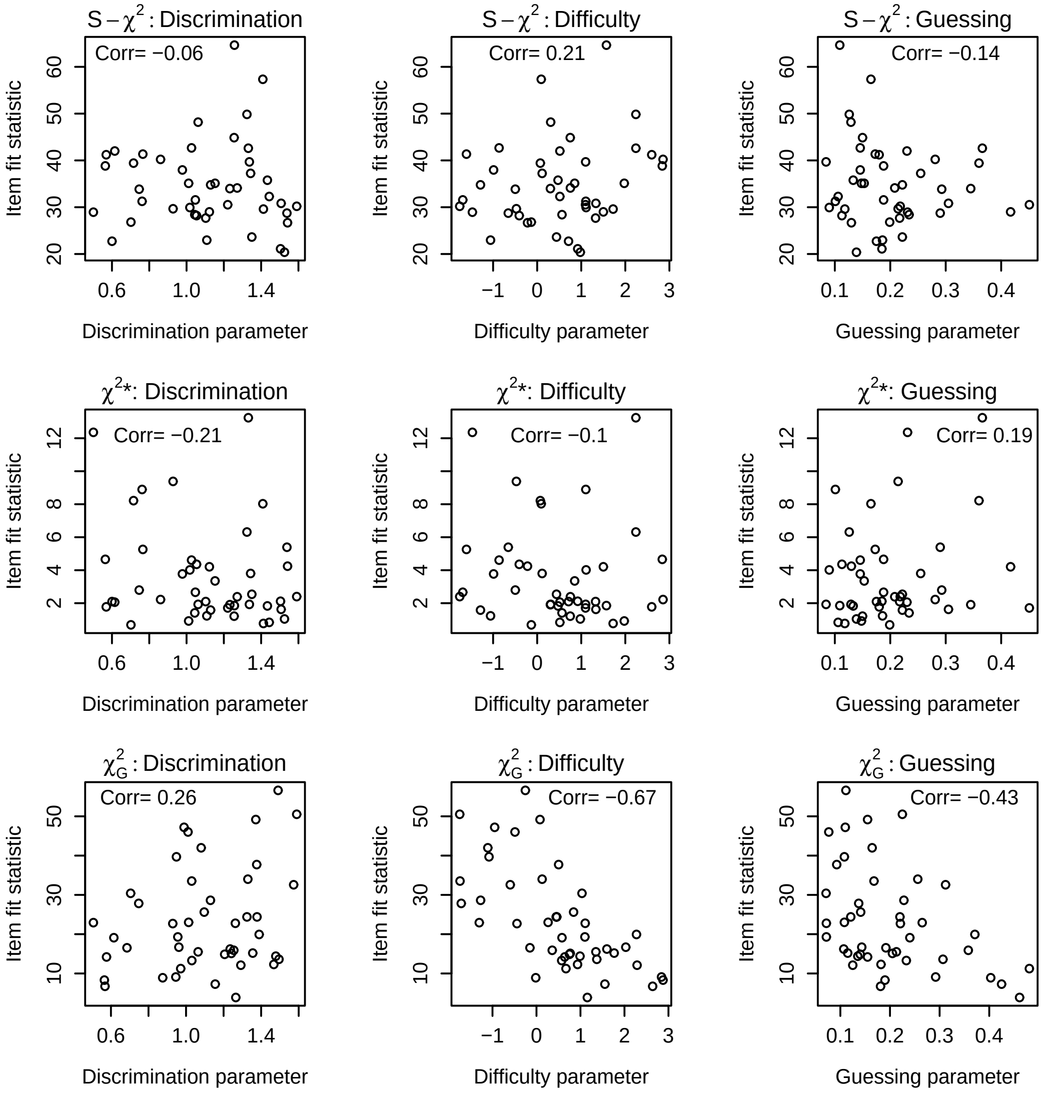

Figure **1 lot of item parameter estimates versus item fit statistics for the real data example.**

A Closer Look at the Statistics Used by Dodeen

Why is the χ^2_G statistic affected by the average slope parameters, while the $S - \chi^2$ statistic is not?

The null distribution of the χ^2_G statistic used in Dodeen (2004) is not χ^2, even asymptotically, as assumed in that paper (p. 264). Dodeen found the Type I error rates at the 1% level of χ^2_G to lie between 9 to 28%. This paper (see Table 3) also found the rate to be much higher than the nominal level, especially for short tests. Stone and Zhang (2003) computed the Type I error rate for the χ^2_B statistic, which is similar to the χ^2_G statistic, for a variety of situations under the 3PL model—the rates are extremely high (sometimes 100%, i.e., every item is labeled as misfitting) for 10 items and 20 items; the rate is much larger than the nominal level even for 40 items and 1,000 or 2,000 examinees (respectively, 0.11 and 0.32 at the 5% level). Orlando and Thissen (2000, 2003) and Glas and Suarez-Falcon (2003), using detailed simulations, found the Q_1 statistic (Yen, 1981), which is also similar to χ^2_G, to have a Type I error rate considerably higher than the nominal level.

The main reasons for the poor behavior of the χ^2_G statistic is that it uses point estimates of ability and ignores the uncertainty in the ability estimates while computing the p-value (see, e.g., Stone, 2000, p. 59). Besides, Chernoff and Lehmann (1953) showed that a χ^2 test statistic computed from numbers of individuals falling into specified cells (the ability groups in this context) does not have a limiting χ^2 distribution when estimates of parameters from the original observations (the item response data in this context) are used. Instead, such a statistic is stochastically larger than what is obtained under χ^2 theory; the departure may be significant for a small number of cells. This is another reason why the χ^2_G statistic is not expected to follow a χ^2 distribution.

In fact, Ansley and Bae (1989) found the Q_1 statistic to have an approximate noncentral χ^2 distribution for the 3PL model in a simulation study—the noncentrality parameter should depend on the parameters of the model in a complicated manner. We anticipate that the χ^2_G statistic behaves like the Q_1 statistic, and the simulations in Dodeen (2004) might have captured some level of the dependence.

On the other hand, detailed simulations under a variety of conditions in this paper and the references mentioned earlier suggest that when data come from a 3PL model, the Type I error level of $S - \chi^2$ approaches the nominal level irrespective of any other factors (including item parameters). Thus, there is no reason to expect any relationships for $S - \chi^2$, as observed for χ^2_G by Dodeen (2004).

Discussion and Recommendations

Dodeen (2004) found some linear relationships between item parameters and item fit statistics in a simulation study. This paper replicates Dodeen's simulations and performs further simulations using test statistics ($S-\chi^2$, $S-G^2$, χ^{2*}, and G^{2*}) that differ in significant ways from those used in Dodeen: (a) these were suggested recently, and (b) each of these statistics has often been found to have a Type I error rate very close to the nominal level. This paper demonstrates that if one uses the $S-\chi^2$ and $S-G^2$ statistics, there is no reason to worry about any linear relationships between the item parameters and item fit statistics when data come from the hypothesized model. This finding will come as a relief to practitioners using these statistics. Interestingly, some linear relationships were found between the item parameters and the χ^{2*} and G^{2*} item fit statistics when data come from the 3PL model; besides, the Type I error rate for the χ^{2*} and G^{2*} were found to be rather high for several test conditions.

Given these findings, a wise option for practitioners will be to use the $S-\chi^2$ and $S-G^2$ statistics. The first principle in statistical hypothesis testing is that a hypothesis is "innocent until proven guilty," and test statistics with Type I error rates higher than the nominal level violate the first principle. Further, the power of the $S-\chi^2$ and $S-G^2$ statistics have been found to be respectably high in several studies. An IRT practitioner using item fit statistics with poor Type I error properties (like χ^2_G and z_j) must be prepared for consequences like those found in Dodeen (2004). For example, our real data example shows a negative correlation between estimated difficulty parameters and χ^2_G values, and also between estimated guessing parameters and χ^2_G values. (Note that these correlations were positive in Dodeen's study.) It is entirely possible that another data set could reveal a relationship that is totally different from what we have shown in this paper. Thus, the poor Type I error rate property of χ^2_G may manifest itself in different ways in different applications. It is true that Dodeen found a factor (item parameters) explaining high Type I error rates for these statistics; however, the levels of correlations found in Dodeen are quite low (the maximum is 0.42) and not enough to describe exactly when the χ^2_G and z_j statistics wrongly show misfit. There is no obvious method for using the findings in Dodeen in some way to obtain a corrected version of χ^2_G whose Type I error rate is close to the nominal level.

One advantage of the statistics considered in Dodeen (2004) is that they are available in a number of standard statistical software packages. However, the GOODFIT software (Orlando, 1997) for computing the $S-\chi^2$ and $S-G^2$ statistics is available for free from Orlando.

Two issues regarding item fit are not covered in this paper and are possible topics for future research. First, this study, as in Dodeen (2004), examines only the linear relationships between item parameters and item fit statistics; a thorough study of the nonlinear relationships between them would be of interest. Also, this study simulated, as in Dodeen, items that should not show any misfit because data were generated under the 3PL model. It would be interesting to perform simulations for cases in which items are supposed to show misfit because they are generated from a model inconsistent with the 3PL model. It may be possible to find interesting relationships between type of item misfit and the values of item fit statistics.

The average values of the statistics used in this paper may not be comparable over test conditions because the degrees of freedom for these statistics are often different over items and replications. We still reported the averages to make our results comparable to those of Dodeen (2004). It is possible to divide the values of the item fit statistics by the corresponding degrees of freedom before averaging to produce an average per degree of freedom value of the fit statistics, and then compare those quantities over test conditions. In such a comparison, the results for χ^2_G and $S - \chi^2$ were found to be the same as those reported earlier (i.e., the average per degree of freedom value increased with an increase in average discrimination parameter for χ^2_G and did not change with an increase in average discrimination parameter for $S - \chi^2$). The results for χ^{2^*} were different from those reported earlier (i.e., the average per degree of freedom value of χ^{2^*} did not increase with an increase in average discrimination parameter; instead, it often decreased). It is also possible to use a different design than the one used in this study. In particular, using predetermined generating item parameters (e.g., as in Section 6 of Sinharay, 2006) may provide further insight, especially about any possible nonlinear relationship between item parameters and item fit statistics.

Though the message of this paper is that item parameters are not linearly associated with the values of certain item fit statistics studied, IRT practitioners would like to know what factors influence item fit statistics. Finding the particular type of content and/or other item characteristics that are likely to result in item misfit will benefit test developers substantially. Such knowledge may be obtained by performing detailed item fit analyses of real data sets, in the same way DIF analyses are performed on real test data to explore factors affecting DIF (see., e.g., Schmitt et al., 1993, and the references therein).

References

Ansley, T. N., & Bae, H. W. (1989, April). *An empirical investigation of the nature of the distribution of an IRT goodness-of-fit statistic.* Paper presented at the annual meeting of the American Educational Research Association, San Fransisco, CA.

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

Chernoff , H., & Lehmann, E. L. (1953). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Annals of Mathematical Statistics, 25*, 579–586.

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

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.

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

Leung, C., Chang, H. H., & Hau, K. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympson-Hetter algorithm. *Applied Psychological Measurement, 26*, 376–392.

Lu, Y. & Lin, S. (2005, April). *Assessing fit of item response theory models.* Poster presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Mislevy, R., & Bock, R. D. (1990). BILOG: Item analysis and test scoring with binary logistic models [Computer software]. Mooresville, IN: Scientific Software.

Orlando, M. (1997). Item fit in the context of item response theory. (Doctoral dissertation, University of North Carolina, 1997). *Dissertation Abstracts International*, 58, 2175.

Orlando, M., & Thissen, D. (2000). Likelihood-based 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.

Rohatgi, V. K. (1976). *An introduction to probability theory and mathematical statistics.* New York: John Wiley.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypothesis about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42*, 375–394.

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

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. (2004). IRTFIT_RESAMPLE: A computer program for assessing goodness of fit of item response. *Applied Psychological Measurement, 28,* 143–144.

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.

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

Notes

[1] Dodeen (2004) mentions (p. 264) that he used the statistic $\chi^2_B = \sum_{j=1}^{n} \frac{N_j(O_j - E_j)^2}{E_j(1-E_j)}$ (Bock, 1972), but also mentions (pp. 264, 266) that BILOG 3.11 (Mislevy & Bock, 1990) was used to compute the statistic. Because BILOG 3.11 computes χ^2_G and not χ^2_B, we assume that Dodeen actually reported results for χ^2_G.