Running head: EQUIVALENCY TEST FOR MODEL FIT

An Equivalency Test for Model Fit

Craig S. Wells

University of Massachusetts – Amherst

James. A. Wollack

Ronald C. Serlin

University of Wisconsin – Madison

Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada, April 12-14, 2005.

Abstract

Yen's Q_1 statistic was adapted to test the fit of the two-parameter logistic model (2PLM) in item response theory (IRT), whereby a significant result implies that the model fits given an acceptable magnitude of misfit. A Monte Carlo simulation study was performed to evaluate the empirical Type I error rate and power of Q_1 under the following crossed conditions: test length (20-, 40-, and 80-items) and sample size (2500-, 5000-, and 10000-examinees). The fit statistic exhibited conservative Type I error rates for larger sample sizes and longer test lengths. The power was adequate in the larger sample size conditions.

An Equivalency Test for Model Fit

Item response theory (IRT) is a model-based measurement theory that provides several attractive features for practitioners, such as invariance of item and person parameters. However, in order for a practitioner to take advantage of IRT's benefits, certain assumptions must be satisfied (e.g., local independence). A common assumption, not often discussed as such, pertains to the fit of the model to the data; i.e., the model being used must accurately portray the true relationship between ability and performance on the item. Model misfit has several consequences for the measurement process, including possibly leading to a violation of the invariance property (see Bolt, 2002). As a result, it is wise for a test developer to establish that a particular model fits the data before operationalizing a valid item.

There are several model misfit statistics available in practice (e.g., BILOG's G^2 and Orlando and Thissen's (2000) $S\text{-}X^2$). For all existing misfit statistics, the null and alternative hypotheses are specified, in general terms, as follows:

$$\begin{aligned} H_0 &: \text{ Model fits the data exactly} \\ H_1 &: \text{ Model does not fit the data exactly.} \end{aligned} \tag{1}$$

Consequently, the model is initially assumed to fit exactly; if the evidence of misfit is substantial, the null hypothesis of fit is rejected, and it is concluded that the model does not fit. There are two main disadvantages to specifying the hypotheses in this manner. First, since the point-null hypothesis, specified in (1), is *always* false (Cohen, 1994), H_0 will inevitably be rejected as the sample size becomes large, even in the presence of a trivial amount of misfit. As a result, a test developer is left interpreting the practical meaningfulness of a significant result. This problem further highlights test developers'

unrealistic expectations that a parametric model will be able to explain the underlying item response function perfectly. One possible solution to this problem is to restate the hypotheses as follows:

$$\begin{aligned} H_0 &: \text{ Model provides close fit to the data} \\ H_1 &: \text{ Model does not provide close fit to the data.} \end{aligned} \tag{2}$$

Notice that the hypotheses specified in (2) no longer require exact fit. As a result, H_0 is not always false, and the aforementioned overpower problem for large sample sizes will not exist.

A second disadvantage of (1), which is also an issue with (2), is that it is not possible to conclude that the model fits the data, because failing to reject H_0 is not the same as accepting H_0 to be true (Tryon, 2001; Rogers, Howard, & Vessey, 1993). However, within the context of model fit, concluding that the model fits the data is precisely the conclusion that one wishes to be able to draw. The purpose of the present study is to describe a new model-fit testing procedure that avoids these two disadvantages. In this study, we present a test for model fit in which H_0 is not always false and which permits the conclusion that the model fits the data.

There are two main types of errors that one can make when performing a hypothesis test: Type I and II. A Type I error is committed when a true H_0 is rejected, whereas a Type II error occurs when a false H_0 is not rejected. Because it is easier to control the probability of committing a Type I error, the null and alternative hypotheses should be specified so that the worse mistake that can be made is associated with a Type I error. In the context of model fit, the worse mistake in most situations is to conclude that the model fits the data when in fact it does not fit the data. Therefore, the correct formulation of the hypotheses is as follows:

$$\begin{aligned} H_0 &: \text{ Model does not provide close fit to the data} \\ H_1 &: \text{ Model provides close fit to the data .} \end{aligned} \tag{3}$$

The form of hypotheses specified in (3) is a type of range-null hypothesis (Serlin & Lapsley, 1993) often referred to as an equivalency hypothesis. The name equivalency comes from biostatistics, where sometimes the goal is to conclude that two experimental drugs performed comparably (Rogers et. al., 1993). The present study will describe a method for testing model fit for the two-parameter logistic model (2PLM) based on a form of Yen's Q_1 statistic.

Fit Statistic in IRT

Yen (1981) proposed a fit statistic based on comparing the observed proportion correct to the model-based prediction for groups of examinees that were grouped based on ability estimates. The statistic, Q_1, is distributed as a chi-square variate, with degrees of freedom equal to the number of groups minus the number of parameters in the model and is computed as follows:

$$Q_1 = \sum_{g=1}^{G} n_g \frac{(p_g - P_g)^2}{P_g(1-P_g)}, \tag{4}$$

where G represents the number of groups; p_g and P_g indicate the observed proportion correct and the model-based prediction for group g, respectively; and n_g denotes the number of examinees in group g.

Unfortunately, Q_1 exhibits inflated Type I error rates because, in practice, examinees can only be grouped based on theta estimates rather than true theta values. Orlando and Thissen (2000) developed a procedure for grouping examinees based on the number correct (NC) score instead of the model-based theta estimate. Using this new procedure for forming groups, Orlando and Thissen (2000) found that their $S\text{-}X^2$ index,

which is structurally identical to Yen's (1981) Q_1 statistic, controlled the Type I error rate. In the present study, a strategy similar to that developed by Orlando and Thissen (2000) for using raw scores to group examinees was used to test the equivalency hypothesis stated in (3). The following discussion details how Yen's Q_1 test statistic was calculated for item *i*.

1. Obtain an estimate of θ for each examinee *j* by converting the rest score (i.e., raw score excluding item *i*) to percentiles ($rank_j/(N+1)$) and then to corresponding *z*-scores.
2. Compute observed proportion correct $\left(p_g\right)$ for item *i* for each unique estimate of θ (i.e., group *g*). Note that the number of *p*'s for item *i* equals the number of unique θ estimates (i.e., raw scores).
3. Calculate the model-based probabilities for each group, P_g, which are defined by the unique θ estimates, by finding the optimal α and β for the 2PLM using a straight forward MLE technique (Bolt, 2002) given the vector of p_g and unique θ estimates.
4. Compute Q_1 in the form described by Yen (1981).

The advantage of this procedure is that it not only takes advantage of Orlando and Thissen's use of grouping examinees based on raw scores, it also lends itself to testing the equivalency hypothesis specified in (3).

Equivalency Testing Procedure

In order to test for equivalency, a test developer must specify a null and alternative hypothesis that states an acceptable amount of misfit that can be tolerated. One such criterion may be specified with respect to how far the true probability of a

correct response is from the model-based estimate for each group. For example, suppose that a test developer is willing to use a model-based estimate for an item as long as its predictions are, on average, within δ of the true probability. Therefore, H_0 and H_1 can be written as follows:

$$\begin{aligned} H_0 &: E(P_g - \pi_g)^2 \geq \delta^2 \\ H_1 &: E(P_g - \pi_g)^2 < \delta^2, \end{aligned} \tag{5}$$

where P_g and π_g represent the model-based and true probability of group g correctly answering an item, respectively.

Once the criterion and null hypothesis have been specified, the next step is to obtain the critical values for testing H_0. In testing the equivalency hypothesis, it is necessary to construct two distributions under H_0: one for an upper and one for a lower bound. As a visual demonstration, consider the item characteristic curve (ICC) for a particular item shown in Figure 1.

Insert Figure 1 about here

The solid line represents the model-based probability (P_g) while the dashed lines represent the upper $\left(\pi_g^U\right)$ and lower $\left(\pi_g^L\right)$ bounds defined by the criterion of $\delta = .025$; i.e., for each value of θ, an error of $\pm .025$ is deemed acceptable.

Each distribution under H_0 is based on a noncentral chi-square with degrees of freedom equal to the number of groups minus the number of item parameters, and a noncentrality parameter (λ), which is calculated for the upper and lower bounds as follows:

$$\lambda_U = \sum_{g=1}^{G} n_g \frac{(\pi_g^U - P_g)^2}{P_g(1-P_g)} \text{ and } \lambda_L = \sum_{g=1}^{G} n_g \frac{(\pi_g^L - P_g)^2}{P_g(1-P_g)} \quad . \tag{6}$$

The critical value for both null distributions is the chi-square value associated with the 5th percentile of the noncentral chi-square, which is in contrast to the 95th percentile representing the critical value of a test of model misfit specified in (1). Therefore, if the observed statistic is less than the 5th percentile in *both* noncentral distributions, we can reject H_0 and conclude that the model provides good-enough fit. Each test can be done using a full α because, for any particular item, P_g cannot simultaneously be above and below π_g by at least δ (i.e., only one of the null distributions can be true).

Method

A Monte Carlo simulation study was performed to assess the empirical Type I error rate and power of the proposed model fit statistic. Dichotomous data were generated under the following crossed conditions: test length (20-, 40-, and 80-items) and sample size (2500-, 5000-, and 10000-examinees). The large sample sizes have been chosen to represent the difficulty that large-scale standardized testing programs face when assessing fit in the presence of "too much" power.

To assess the Type I error rate (which, under the frameworks in (3) and (5), is the percentage of misfitting items identified as fitting within some pre-specified criterion), 20% of the items were simulated from the three-parameter logistic model (3PLM) such that the magnitude of misfit was slightly beyond the criterion defined under H_0 (i.e., unacceptable amount of misfit). The magnitude of misfit was measured as follows:

$$MISFIT = \sqrt{\sum_{j=1}^{N} w(\theta_j)\left(P_{3PLM,j} - P_{2PLM,j}\right)^2} \quad . \tag{7}$$

$P_{3PLM,j}$ represents the generating probability given by the 3PLM, while $P_{2PLM,j}$ represents the probability of the best-fitting 2PLM as obtained via an MLE technique (Bolt, 2002). $w(\theta_j)$ represents the weight, defined by the normalized density of the standard normal, for a particular θ_j value. For purposes of assessing the Type I error rate, generating item parameters that produced *MISFIT* values of .025 (i.e., $\delta = .025$) were used to simulate misfit.

To assess the power of the model fit statistic to detect items for which the 2PLM fit adequately, 40% of the items were simulated from the 2PLM while the remaining 40% were simulated from the 3PLM that corresponded to *MISFIT* values of around .01. The generating parameter values are reported in Table 1. The first 20% of the items of a particular test length (e.g., items 1-16 for test length of 80 items) correspond to the parameter values used to simulate misfit while the following 40% correspond to the "fitting" 3PLM parameter values. The remaining set of parameter values belong to the 2PLM and were selected from an English placement examination used in a Midwestern university system. In addition, the generating parameter values were hierarachically nested within test length, so that the 40-item test consisted of half of the items on the 80-item test and the 20-item test consisted of half of the items on the 40-item test.

Insert Table 1 about here

The underlying theta values were sampled from the standard normal distribution, $\theta \sim N(0,1)$. 1000 replications were performed for each of the 9 conditions (test length X sample size).

Code within the software package R was used to compute the fit statistic, Q_1, and to examine the empirical Type I error rate and power for each item at an α level of .05.

Results

Empirical Type I Error Rate

Table 2 reports the detection rate for each condition averaged across items that were simulated to exhibit misfit slightly beyond the acceptable criterion of $\delta = .025$.

Insert Table 2 about here

Although the overall empirical Type I error rate was acceptable, there were apparent sample size and test length effects. As the sample size increased, the empirical Type I error rate decreased noticeably, especially for the 40- and 80-item test length conditions. As test length increased from 20 to 40 items, the Type I error rate dropped.

Empirical Power

Table 3 reports the detection rate for each condition averaged across items that were simulated from the 2PLM and 3PLM with acceptable misfit (i.e., items 17-80 in Table 1).

Insert Table 3 about here

There were apparent sample size and test length effects in that the probability of detecting fitting items increased with larger sample sizes but decreased with longer test lengths. Furthermore, the detection rate was larger for items generated from the 2PLM versus those generated from the 3PLM.

Discussion

The fit procedure described in this paper was developed to test a null hypothesis whose rejection would imply that the model fits, given an acceptable discrepancy between the average true probability of a correct response and model-based prediction. The advantage of such a method is that it tests a meaningful hypothesis and is particularly useful in evaluating fit for large samples in which the traditional point-null hypothesis is usually rejected, even for a trivial amount of misfit. Although the procedure was conservative for larger sample sizes and test lengths, it did not exhibit an inflated Type I error rate and was able to detect fitting items at an adequate level in the larger sample size conditions.

One possible explanation for why the statistic exhibited a conservative Type I error rate for larger sample sizes may be due in part to how the misfit was simulated. To assess the Type I error rate at a specific α level, it is important to simulate misfit exactly at the criterion specified under H_0. Unfortunately, it is difficult, if not impossible, to select various, realistic generating item parameter values for the 3PLM that simulate the misfit at exactly the specified δ value. As a result, generating item parameter values were chosen that produced *MISFIT* values ranging from very close to slightly beyond .025. Therefore, as the sample size increased, it was easier to identify the item as misfitting, resulting in a lower probability of rejecting the null hypothesis that the item

did not fit. Hence, the further the *MISFIT* values are from the criterion, the lower the Type I error rate will be, particularly for large sample sizes where item parameters are estimated with less error.

A critical element of the range-null hypothesis approach is specifying the particular value for δ. For example, the overall power of the method is affected partly by the specified δ value, with larger values leading to increased power, presuming a non-inflated Type I error rate. Although the particular value of δ selected for this study may be considered reasonable (albeit perhaps a little small), it is by no means the only sensible value, and in fact, was chosen more for illustrative purposes than as a suggestion for practitioners. It should be clear that specifying the criterion under H_0 is not trivial and requires further research before acceptable values may be given as guidelines. When performing such research, it is most important to consider how misfit influences the validity of an IRT scale through loss of the invariance property (Bolt, 2002). Therefore, misfit may have a similar effect on the measurement process as differential item functioning (DIF); e.g., it may corrupt the scale through its effect on the equating process under certain conditions (Shepard, Camilli, & Williams, 1984).

Although only the 3PLM was used to simulate item responses for which the 2PLM either "fit" or misfit, depending on the average discrepancy between the true IRF and optimal 2PLM IRF, it is certainly not the only model that could have been used to simulate non-2PLM IRFs. In fact, various forms of misfit are conceivable when applying parametric IRT models. Another type of non-2PLM IRF that may occur is where the true IRF departs from the typical logistic, S-shaped curve. Such an IRF may be generated using the mixture nominal response model (MNRM; Bolt, Cohen, & Wollack, 2001).

Interestingly, whereas the 3PLM tends to produce misfit at the extremes of the ability distribution where the lower asymptote is above zero, the 2PLM misfits the IRF produced by the MNRM in the middle of the ability distribution. Therefore, using different models to simulate misfit allows one to explore the impact that different types of misfit have on the fit statistic.

An advantage of the method described herein is that it can be easily extended to test the fit of other parametric models such as the 3PLM, or even models for polytomous data such as Samejima's Graded Response Model. Furthermore, it will be worthwhile to consider applying the range-null approach to other available statistics, such as Orlando and Thissen's S-X^2 and Glas and Falcon's LM test (2003).

References

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

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26*, 381-409.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

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

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

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*, 553-565.

Serlin, R.C. & Lapsley, D.K. (1993). Rationality in psychological research; The good-enough principle. *American Psychologist, 40*, 73-83.

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

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

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*, 371-386.

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

Table 1. *Generating item parameter values.*

	80-item			40-item			20-item		
Item	α	β	γ	α	β	γ	α	β	γ
1	1.53	-0.31	0.33	1.53	-0.31	0.33	1.53	-0.31	0.33
2	1.19	0.35	0.31	1.19	0.35	0.31	1.19	0.35	0.31
3	1.06	1.43	0.25	1.06	1.43	0.25	1.06	1.43	0.25
4	1.20	0.57	0.23	1.20	0.57	0.23	1.20	0.57	0.23
5	1.28	0.03	0.41	1.28	0.03	0.41			
6	1.29	0.22	0.26	1.29	0.22	0.26			
7	1.11	0.76	0.28	1.11	0.76	0.28			
8	1.08	1.14	0.26	1.08	1.14	0.26			
9	1.11	1.16	0.22						
10	1.08	0.98	0.42						
11	1.07	1.57	0.31						
12	1.10	1.27	0.21						
13	1.08	1.15	0.43						
14	1.07	1.38	0.37						
15	1.33	-0.07	0.42						
16	1.34	-0.08	0.38						
17	0.82	0.46	0.16	0.82	0.46	0.16	0.82	0.46	0.16
18	0.74	0.19	0.29	0.74	0.19	0.29	0.74	0.19	0.29
19	0.75	0.59	0.19	0.75	0.59	0.19	0.75	0.59	0.19
20	0.94	0.55	0.10	0.94	0.55	0.10	0.94	0.55	0.10
21	0.76	0.48	0.20	0.76	0.48	0.20	0.76	0.48	0.20
22	0.87	0.51	0.13	0.87	0.51	0.13	0.87	0.51	0.13
23	0.97	-0.62	0.29	0.97	-0.62	0.29	0.97	-0.62	0.29
24	1.05	-0.47	0.19	1.05	-0.47	0.19	1.05	-0.47	0.19
25	0.99	0.01	0.14	0.99	0.01	0.14			
26	0.85	0.08	0.19	0.85	0.08	0.19			
27	1.38	-0.86	0.19	1.38	-0.86	0.19			
28	0.82	1.17	0.10	0.82	1.17	0.10			
29	0.82	-0.14	0.27	0.82	-0.14	0.27			
30	1.22	0.24	0.07	1.22	0.24	0.07			
31	0.78	0.58	0.17	0.78	0.58	0.17			
32	1.10	-0.74	0.25	1.10	-0.74	0.25			
33	1.05	1.25	0.52						
34	0.87	0.26	0.15						
35	1.04	0.12	0.11						
36	1.19	0.06	0.87						
37	0.85	0.58	0.13						
38	1.31	-0.70	0.16						
39	0.85	0.34	0.15						
40	0.94	1.54	0.06						
41	0.86	1.62	0.08						
42	0.84	0.51	0.14						

43	1.18	-0.01	0.10				
44	0.81	0.56	0.15				
45	0.89	0.01	0.17				
46	1.08	-0.42	0.17				
47	1.14	0.66	0.06				
48	0.83	0.17	0.19				
49	1.25	-1.13		1.25	-1.13	1.25	-1.13
50	0.79	-0.21		0.79	-0.21	0.79	-0.21
51	0.99	-0.37		0.99	-0.37	0.99	-0.37
52	1.23	0.65		1.23	0.65	1.23	0.65
53	0.98	0.02		0.98	0.02	0.98	0.02
54	0.99	-0.17		0.99	-0.17	0.99	-0.17
55	1.66	-0.18		1.66	-0.18	1.66	-0.18
56	1.22	1.28		1.22	1.28	1.22	1.28
57	1.03	-0.95		1.03	-0.95		
58	0.74	0.54		0.74	0.54		
59	1.47	-1.01		1.47	-1.01		
60	0.88	-0.60		0.88	-0.60		
61	1.01	-0.48		1.01	-0.48		
62	0.85	0.55		0.85	0.55		
63	0.73	0.03		0.73	0.03		
64	1.10	-1.26		1.10	-1.26		
65	1.66	-0.18					
66	0.95	-0.38					
67	1.25	-0.53					
68	0.87	0.70					
69	0.83	-0.29					
70	1.23	0.65					
71	0.73	0.03					
72	1.20	-1.68					
73	1.15	-1.21					
74	1.07	-0.74					
75	1.26	-1.06					
76	0.71	-0.51					
77	0.48	1.18					
78	1.23	-0.34					
79	1.17	0.16					
80	1.15	-1.18					

Table 2. *The empirical Type I error rate for each condition averaged across items that were simulated to misfit slightly beyond the acceptable criterion of* δ *=.025.*

		Test Length		
		20-item	40-item	80-item
Sample Size	2500-examinee	.061	.054	.051
	5000-examinee	.055	.042	.044
	10000-examinee	.049	.026	.021

Table 3. *The detection rate for each condition averaged across items that were simulated from the 2PLM and "fitting" 3PLM.*

		Test Length					
		20-item		40-item		80-item	
	Generating Model	2PLM	3PLM	2PLM	3PLM	2PLM	3PLM
Sample Size	2500-examinee	.261	.155	.204	.149	.139	.116
	5000-examinee	.573	.317	.514	.300	.372	.252
	10000-examinee	.894	.590	.886	.574	.809	.506

Figure 1. An ICC with $\alpha = 1.37$ and $\beta = -0.21$, along with the upper and lower bounds of acceptable misfit based on $\delta = .025$.