Comparing the Two- and Three-Parameter Logistic Models via Likelihood Ratio Tests: A Commonly Misunderstood Problem

Applied Psychological Measurement

2015, Vol. 39(5) 335–348

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621614563326

apm.sagepub.com

Christy Brown¹, Jonathan Templin², and Allan Cohen³

Abstract

Selection of an appropriate item response model is critical in the measurement of latent examinee ability. The one-, two-, and three-parameter logistic (1PL, 2PL, and 3PL) models are nested, and as such can be compared using likelihood ratio (LR) tests. The null hypothesis in the LR test for selection among the 2PL and 3PL models sets the guessing parameters to their lower bound of 0. This violates one of the assumptions of the LR test and renders the usual χ^2 reference distribution inappropriate for the comparison. A review of the current literature revealed that this problem is not well understood in the educational measurement field. Ignoring this issue can lead to selection of an overly simplified model, with implications for the ability estimates. In this article, the use of the LR test for item response model selection is investigated, with the goal of providing practitioners with an appropriate method of selecting the most parsimonious model. The results of simulation studies indicate the nature of the problem, with inaccurate Type I error rates for cases where the inappropriate null distribution was used. An analysis of data from a statewide mathematics test showed differences pertinent to subsequent analyses.

Keywords

item response theory, model selection, likelihood ratio test, boundary issue

Item response models provide a means for measuring the latent ability of each examinee on a set of test items. The estimate of an examinee's ability depends on both the examinee's responses to the set of items and on the properties of these items, as characterized by the statistical parameters used in the selected item response model. The most commonly used unidimensional item response theory (IRT) models for dichotomously scored items include the one-parameter logistic (1PL) model, the two-parameter logistic (2PL) model, and the

¹Clemson University, Clemson, SC, USA

²The University of Kansas, Lawrence, KS, USA

³The University of Georgia, Athens, GA, USA

Corresponding Author:

Christy Brown, Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, USA.

Email: cjb2@clemson.edu

three-parameter logistic (3PL) model (e.g., Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Lord & Novick, 1968). For researchers and analysts intending to measure examinee latent traits, selection of an appropriate item response model is critical, the goal being to find the simplest model that adequately explains the observed item responses (i.e., the most parsimonious model). When maximum-likelihood estimation is used, the likelihood ratio (LR) test can be used to select the most parsimonious among the 1PL, 2PL, and 3PL item response models as they form a series of nested models (Waller, 1981). In this article, the statistical properties of the LR test for model selection in unidimensional IRT models are investigated.

The authors of the present study begin by defining the IRT models of interest and illustrating their hierarchical nature. Next, the authors discuss how the LR testing procedure fits into an overall assessment of model-data fit in the IRT framework. They then describe how LR tests involving the 3PL model violate one of the assumptions underlying the test, a fact that has gone unrecognized in the IRT literature. Next, several simulation studies illustrating the impact of ignoring the assumptions violation on the Type I error rate and power of LR tests comparing the 2PL and 3PL models are presented. Then, a method of accounting for these nonstandard conditions is recommended when conducting these tests in practice, which is illustrated in an analysis of data from a statewide mathematics test. Finally, some discussion of the results with an emphasis on the potential consequences of ignoring the boundary issue when using LR tests to select an IRT model is provided.

The 1PL, 2PL, and 3PL Item Response Models

In the 3PL model, the probability that examinee i answers item j correctly for $i = 1, \dots, N$ and $j = 1, \dots, n$ is given as follows:

$$P_j(\theta_i) = P(Y_{ij} = 1 | \theta_i) = c_j + (1 - c_j) \frac{\exp\{a_j(\theta_i - b_j)\}}{1 + \exp\{a_j(\theta_i - b_j)\}}, \quad (1)$$

where Y_{ij} is the dichotomously scored response of person i to item j ($1 = \text{correct}$, $0 = \text{incorrect}$), θ_i is the latent ability of person i , a_j is the discrimination of item j , b_j is the difficulty of item j , and c_j is the guessing parameter (or lower asymptote) for item j with $0 \leq c_j \leq 1$. The 2PL model is nested within the 3PL model, as it can be obtained by constraining the guessing parameters $c_j = 0$ for all j . Similarly, the 1PL model is nested within the 2PL and 3PL models because it can be obtained from the 2PL model by setting all item discrimination parameters equal to a common value, $a_j = a$. A special case of the 1PL model is the Rasch model, in which $a_j = 1$ for all j .

A common approach to estimating the parameters in these IRT models is marginal maximum-likelihood (MML) estimation (see, for example, Bock & Aitkin, 1981). The popular IRT software programs BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) and MULTILOG (Thissen, 1991) both implement MML estimation of item parameters when no prior distributions are specified, and by default calculate Bayesian estimates of the latent ability parameters.

IRT Model Fit Evaluation

In evaluating the goodness of fit of a statistical model to sample data, a distinction can be made between assessing *absolute* model fit and assessing *relative* model fit. Absolute model fit is concerned with how closely responses predicted by the model align with the responses that were observed, whereas relative model fit is concerned with the selection of the model that best

fits a data set efficiently from a set of candidate models. Several methods for evaluating the absolute fit of IRT models have been proposed. Many of these techniques are inference based (e.g., Andersen, 1973; Bock, 1972; Glas, 1999; Orlando & Thissen, 2000; Yen, 1981, 1984). Others have proposed methods incorporating graphical techniques and the examination of residuals in assessing the absolute fit of IRT models (see, for example, Hambleton, Swaminathan, & Rogers, 1991).

In evaluating relative model fit, the goal is to find the simplest model in a set of candidate models that adequately explains the observed responses. The Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978) are two measures of relative fit that account for model complexity in their selection of the best fitting model, but are not inferential procedures. However, statistical significance tests do exist for measures of relative fit when the competing models are nested, as is the case for the 1PL, 2PL, and 3PL models. One such statistical significance test is the LR test, which is the focus of this article. See Kang and Cohen (2007) for a comparison of the performance of various measures of relative model fit, including AIC, BIC, and the LR test for selection of an IRT model. In using measures of relative fit, there is no guarantee that the selected model fits the data well, only evidence that it fits at least as well as the larger models considered. Hence, it is essential that tests of absolute fit be used in conjunction with those of relative fit. The study of Maydeu-Olivares and Cai (2006) demonstrates this point in the context of IRT models.

For many large-scale assessments, such as state end-of-course tests, policy often dictates the choice of IRT model. Such programs are then interested solely in the absolute fit of the chosen IRT model to observed responses, and measures of relative fit are not of concern. However, LR tests for IRT model selection have been used with smaller-scale assessments in studies from a variety of contexts, including education research (Bergan, 2010), personality data evaluation (Reise & Waller, 2003), language testing research (Bachman, Davidson, Ryan, & Choi, 1995), psychological testing research (Kline, 2005), and health outcomes instrument development (Edelen & Reeve, 2007). Such studies rely on the description of how to implement LR tests in IRT model selection provided in introductory texts on IRT (e.g., De Ayala, 2009; DeMars, 2010; Embretson & Reise, 2000) and in the documentation for IRT software programs (e.g., du Toit, 2003; Thissen, 1991). Hence, it is important that the statistical properties of LR tests for comparing IRT models are well understood and properly communicated in such sources. The next section of this article presents the technical details underlying LR tests for IRT model selection, including a discussion of which are nonstandard testing situations.

LR Tests

An LR test assesses whether or not reduction from a more complex or “full” model to a simpler or “reduced” model is appropriate when the two models are nested. Let $\hat{\ell}_R$ denote the maximized log-likelihood of the reduced model and $\hat{\ell}_F$ the maximized log-likelihood of the full model. Then, the LR test statistic is given as follows:

$$G^2 = -2(\hat{\ell}_R - \hat{\ell}_F). \quad (2)$$

Under certain regularity conditions, G^2 has an asymptotic χ^2 distribution with degrees of freedom equal to the difference between the number of parameters in the full and reduced models (Wilks, 1938). The larger the obtained value of G^2 , the stronger the evidence that the reduced model is inadequate and that the full model should be retained.

The LR testing procedure can be applied to the 1PL, 2PL, and 3PL models, as they form a series of nested models. Models with more parameters will always fit at least as well as models with fewer parameters when the models are nested, but the LR test can determine if removal of certain parameters does not significantly decrease the model fit. Assuming measures of absolute model fit indicate that the 3PL model is suitable, an LR test can be used to determine if reduction to the 2PL model is appropriate. If the 2PL model holds, then a second LR test can be used to determine if further reduction to the 1PL model is permissible. Calculation of the LR statistic for such tests is possible in BILOG-MG as well as MULTILOG, as both programs report $-2\hat{\ell}$ values for estimated models in the standard output. In addition to selection of an overall model, LR tests can also be used for model selection at the item level. That is, the most parsimonious model for a set of item response data might be one in which a 3PL model is used for some items and a 2PL for others. MULTILOG can be used to calculate the LR statistic for such comparisons, as its options allow the user to specify different models for each item.

Across all items, the hypotheses in an LR test comparing the 2PL and 3PL models are as follows:

$$H_0: c_j = 0 \text{ for all } j = 1, \dots, n$$

$$H_A: c_j \neq 0 \text{ for at least one } j = 1, \dots, n.$$

The null hypothesis places the guessing parameters at their lower boundary of 0, thereby violating one of the regularity conditions needed for Wilks's theorem to hold. As a result, the standard $\chi^2(n)$ reference distribution is no longer valid.

The problem of testing a null hypothesis that is on the boundary of the parameter space is a familiar one in the mixed-effects modeling literature. Using LR methods to test for the inclusion of random effects in a linear model is a nonstandard problem, as the null hypothesis that their variance components are 0 places these variance parameters on the boundary of the parameter space (Stram & Lee, 1994). In the most common situation, testing for the existence of a single random effect is a test of $H_0: \sigma^2 = 0$ against $H_A: \sigma^2 > 0$. In this case, the limiting distribution of the LR statistic is known to be a 50:50 mixture of the $\chi^2(0)$ and $\chi^2(1)$ distributions, where $\chi^2(0)$ represents a distribution with all of its mass or probability at 0 (Self & Liang, 1987; Stram & Lee, 1994). In the context of IRT models, LR tests where the null hypothesis sets a guessing parameter at its boundary of 0 should also incorporate a nonstandard reference distribution. This, unfortunately, is an issue that has consistently been overlooked when the LR testing procedure for model comparison is described in the IRT literature (e.g., De Ayala, 2009; DeMars, 2010; du Toit, 2003; Embretson & Reise, 2000; Thissen, 1991). To help inform researchers and analysts using LR tests to select an IRT model, next a series of simulation studies designed to illustrate how using the incorrect null distribution in these tests affects IRT model choice are presented.

Simulation Studies

The simulation studies consisted of two main components. In the first component, Type I error rates were estimated for three different LR tests for IRT model selection through simulation of the null distribution. In the second component, the power of these tests to detect varying levels of departure from the null model was investigated across different sample sizes.

Type I Error Studies

Method. The purpose of this section is to illustrate the impact of the boundary condition violation on the Type I error rate in LR tests involving the 3PL model. Type I error rates were calculated in three different studies comprising the following testing scenarios:

1. The first Type I error study considered an LR test for selection among the Rasch and the 2PL models. That is, the hypotheses of interest were $H_0 : a_j = 1$ for all items $j = 1, \dots, n$ versus $H_A : a_j \neq 1$ for at least one j . *In this condition, the boundary issue is not a concern and thus the standard $\chi^2(n)$ reference distribution should hold.*
2. The LR test investigated in the second Type I error study compared the 2PL and 3PL models for all items of a test. Thus, the hypotheses under consideration in this study were $H_0 : c_j = 0$ for all items $j = 1, \dots, n$ versus $H_A : c_j \neq 0$ for at least one j . *The boundary issue is therefore a concern in this case, and this article seeks to demonstrate that the standard $\chi^2(n)$ reference distribution does not hold.* The correct reference distribution in LR tests where the null hypothesis places more than one parameter at a boundary value has been derived only in certain special cases, none of which apply here (for further details, see Silvapulle & Sen, 2005).
3. In the third Type I error study, LR tests for model selection at the item level rather than at the overall model level were considered. The reduced model in this case always specified the 2PL model for all n items. This was tested against a full model that specified the 3PL model for just one of the test items, while the remaining $n - 1$ items were still modeled with the 2PL. Hence, this study consisted of as many LR tests as there were test items, each one testing $H_0 : c_j = 0$ versus $H_A : c_j \neq 0$ for some $j = 1, \dots, n$. The null hypothesis in each of these LR tests placed exactly one parameter at a boundary value, a case for which Self and Liang (1987) showed that the distribution of the LR statistic is a 50:50 mixture of $\chi^2(0)$ and $\chi^2(1)$. *Thus, the boundary issue is also a concern in this case, and the authors demonstrate that the standard $\chi^2(1)$ reference distribution does not hold and that the correct reference distribution is instead a 50:50 mixture of the $\chi^2(0)$ and $\chi^2(1)$ distributions.*

For each of these three scenarios, the distribution of the LR statistic under the null hypothesis was approximated through simulation. This was done to estimate the probability of mistakenly rejecting a true null hypothesis (i.e., the Type I error rate) when the boundary issue is ignored and the standard reference distribution is used to assess significance. In each condition, this was accomplished by generating 1,000 binary item response data sets consisting of $N = 5,000$ examinees and $n = 30$ items according to the null model. The LR test statistic G^2 (see Equation 2) was evaluated in each of the 1,000 replications by estimating the parameters of both the null and alternative models to obtain $-2\hat{\ell}_R$ and $-2\hat{\ell}_F$ for each simulated data set. In all cases, the item parameters were estimated using MULTILOG. In each of the three scenarios, a histogram of the LR test statistics from the 1,000 replications was then used to depict an approximation of the correct reference distribution. For each of the LR tests considered, the empirical probability of making a Type I error when the boundary issue is ignored was calculated as the number of obtained G^2 values exceeding the upper α critical value from the standard χ^2 distribution for the LR test of interest. For the third Type I error study, the empirical probability of making a Type I error rate when significance is assessed according to the nonstandard mixture χ^2 distribution was also calculated for comparative purposes. For all studies, Type I error rates were evaluated at three different nominal α levels: .01, .05, and .10.

Figure 1. Histograms of the simulated distribution of the likelihood ratio statistic under the null hypothesis in each of the Type I error studies.

Note. In all three graphs, the solid line is the density curve of the standard χ^2 reference distribution. For the third Type I error study, the dashed line is the density curve of a 50:50 mixture of χ^2 distributions with 0 and 1 degrees of freedom. 2PL = two-parameter logistic; 3PL = three-parameter logistic.

Examinee abilities in all three conditions were drawn from a standard normal distribution. In the first Type I error study, the data were generated according to a Rasch model with difficulty parameters b_j ranging from -2.25 to 2.25 increasing in increments of 0.15 , with 0 excluded. In the second and third Type I error studies, the 2PL was the generating model. For these conditions, 3 values of the item discrimination parameters a_j (0.75 , 1.25 , 1.75) were crossed with 10 values of the difficulty parameters b_j (-2.5 , -2 , -1.5 , -1 , -0.5 , 0.5 , 1 , 1.5 , 2 , 2.5) to define the item parameters of the generating model.

Results. Histograms of the simulated null distributions in each of the Type I error studies are presented in Figure 1. In each case, the density curve for the standard reference distribution was superimposed on the graph. The simulated null distribution for the first Type I error study appears to be well approximated by the standard $\chi^2(30)$ reference distribution. This is as expected, as the null hypothesis is not testing a parameter at its boundary in LR tests comparing the Rasch and 2PL models. In contrast, it is evident from Figure 1 that the standard $\chi^2(30)$ reference distribution does not fit the simulated null distribution from the LR test comparing the 2PL and 3PL models for a 30-item exam. There is a much higher concentration of values at the lower end of the simulated null distribution than would be expected according to the standard reference distribution. Figure 1 also depicts the simulated null distribution for one of the 30 LR tests from the third Type I error study. Specifically, it is the null distribution for the LR test comparing the 2PL and 3PL models for Item 30, which was specified as the most difficult and highly discriminating item of the exam, and for which rejection of the 2PL model in favor of the 3PL makes the most sense substantively. However, Figure 1 demonstrates that this simulated null distribution was not well approximated by the standard $\chi^2(1)$ reference distribution, as there was a higher concentration of values near 0 than would be expected in a $\chi^2(1)$ distribution. The density curve for a 50:50 mixture of the $\chi^2(0)$ and $\chi^2(1)$ distributions, also superimposed on the graph in Figure 1, appears to be a much better fit to this simulated null distribution

than the standard reference distribution. This is again an expected result, as the null hypothesis of interest in this study placed only one parameter (c_{30}) at its boundary value of 0.

In the first study, the observed Type I error rates were quite close to the nominal level with $\chi^2(30)$ as the reference distribution: for $\alpha = .01$, the observed Type I error rate was .01; for $\alpha = .05$, the observed rate was .06; and for $\alpha = .10$, the observed rate was .11. However, in the second study, none of the LR test results were significant even at the $\alpha = .10$ level according to the $\chi^2(30)$ reference distribution. These estimates support what was observed in the histograms of the simulated null distributions, the standard reference distribution is appropriate for comparing the Rasch and 2PL models but not for comparing the 2PL and 3PL models.

In Table A1 (see the online appendix), empirical estimates of the Type I error rate according to the standard $\chi^2(1)$ reference distribution are compared with the rates according to the correct mixture χ^2 reference distribution for the LR tests of the third Type I error study. Across all items, each with varying levels of specified difficulty and discrimination, the observed Type I error rates according to the 50:50 mixture of χ^2 distributions with 0 and 1 degrees of freedom were consistently close to the nominal α level, whereas the rates according to the $\chi^2(1)$ distribution were consistently below the nominal α level, with the differences becoming more notable as the value of α increases. The results of this third study support the authors' conjecture that LR tests for IRT model selection in which the null hypothesis places only one parameter at its boundary should assess significance according to a 50:50 mixture of the $\chi^2(0)$ and $\chi^2(1)$ distributions.

In the third Type I error study, for each simulated data set 30 different significance tests were conducted, one to determine whether including a guessing parameter was advantageous for each item of the test. Thus, the *familywise error rate*, the probability of rejecting at least one null hypothesis in a series of tests when all of the null hypotheses are true (see, for example, Oehlert, 2000), is of concern. The *comparisonwise error rates* reported in Table A1 are simply the probabilities of rejecting a particular null hypothesis, given that it is true. In conducting a large number of tests, the familywise error rate can become severely inflated if no multiplicity control is used. For example, in 776 of the 1,000 replications in the third Type I error study, at least one of the 30 tested null hypotheses was incorrectly rejected when the correct mixture χ^2 distribution was used to assess significance at a comparisonwise error rate of .05, leading to an estimated familywise error rate of .776. Thus, in practical applications it is important to consider some sort of multiplicity correction such as the Bonferroni procedure, which is quite popular in part because it is easy to implement. For the Bonferroni procedure, if there are K null hypotheses to be tested and the desired familywise error rate is α , then individual tests should be conducted at the α/K level. Applying the Bonferroni procedure to the third Type I error study leads to an estimated familywise error rate of .049, as in only 49 of the 1,000 replications at least one of the 30 tested null hypotheses was incorrectly rejected when the correct mixture χ^2 distribution was used to assess significance at a comparisonwise error rate of $.05/30 = .001\bar{6}$. See Oehlert (2000) for a description of other viable multiple testing correction procedures.

Power Analysis

The results from the Type I error studies illustrate the conservative nature of LR tests comparing the 2PL and 3PL models when the boundary issue is ignored. That is, the probability that these tests will mistakenly reject the simpler 2PL model in favor of the more complex 3PL model is lower than the stated α level. The major drawback of using a conservative test is that it will have less power to detect departures from the null model when they truly do exist. In conducting the following power analysis, the goal was to demonstrate the loss of statistical power that occurs in

LR tests comparing the 2PL and 3PL models when the boundary condition is not taken into account.

Method. The power analysis focused on tests comparing the 2PL and 3PL models for just one item of a simulated 30-item test, as the correct reference distribution in this case is known, thus facilitating a comparison between tests that do and do not account for the boundary issue. The correct reference distribution in an LR test comparing a 2PL model for all items of a test with a 3PL model for all items is not known, and though it is the most common comparison made in practice, the same outcome can be reached using a one-at-a-time approach.

For each condition in this study, response data for 1,000 samples were generated from the alternative (3PL) model to estimate the probability of correctly rejecting a false null hypothesis (i.e., the power of the test). All test lengths were fixed at 30 items. For the generating model, five values of the item guessing parameters c_j (0.05, 0.1, 0.15, 0.2, 0.25) were crossed with two values of the item discrimination parameters a_j (0.75, 1.5) and three values of the item difficulty parameters b_j (-1.5, 0, 1.5). The LR tests considered compared the fit of the alternative model (i.e., the generating 3PL model) with the fit of a null model specifying a 2PL for just one of the items and the correct 3PL model for all other items. Thus, model misspecification was restricted to just a single item when estimating the power of these tests. As the power of a statistical test is dependent upon sample size, samples of 1,000, 2,500, and 5,000 examinees were included.

For each simulated data set, the value of the LR statistic G^2 was determined by estimating the parameters of the alternative (generating) model and the 30 different null models considered using MML estimation as implemented in MULTILOG. For each testing condition, the proportion of observed G^2 values exceeding the upper α critical value for a 50:50 mixture of $\chi^2(0)$ and $\chi^2(1)$ distributions was calculated and compared with the proportion of significant outcomes when the incorrect standard $\chi^2(1)$ distribution was used to assess significance. Two different significance levels will be reported: (a) $\alpha = .05$, which corresponds to no correction for multiple testing, and (b) $\alpha = .001\bar{6}$, which corresponds to a Bonferroni correction to control the family-wise error rate at .05 for the 30 different tests conducted per simulated data set.

Results. The proportion of significant outcomes for a significance level of .05 is given in Table 1. These results would be of interest when only one particular item is being considered for reduction from a 3PL to a 2PL, perhaps because it is known in advance that the item is very easy and it is not thought that a large proportion of examinees will attempt to guess the correct answer choice. As can be seen in Table 1, the proportion of rejections of the misspecified null model increased as the sample size increased across all item parameter specifications. This is an expected result, as larger samples provide more information about the true nature of the population. It is also clear from Table 1 that the values of the difficulty, discrimination, and guessing parameters all greatly affect the power of these tests. It was anticipated that the further the true values of the guessing parameters were from 0, the more likely it would be that a departure from the null value would be detected. However, these tests had adequate power (e.g., higher than 80%) to detect the presence of a guessing parameter only for items that were highly discriminating and at least moderately difficult, even when the value of the guessing parameter was quite large. The degree to which the values of the difficulty and discrimination parameters affected the power estimates is a very interesting and unexpected result.

Also evident in Table 1 is the loss of power that occurred when the significance of an outcome was assessed according to the incorrect standard $\chi^2(1)$ reference distribution as opposed to the correct mixture χ^2 distribution. This was most noticeable for items that were moderately discriminating and at least moderately difficult, as tests for items that were both highly discriminating and highly difficult were quite powerful when either reference distribution was used. For example, consider Items 29 and 30, both with guessing parameters of 0.25 and difficulty

Table 1. Proportion of Rejections of the Misspecified 2PL Model in LR Tests Comparing the 2PL and 3PL Models for Each Item of a 30-Item Test.

Item	<i>c</i>	<i>b</i>	<i>a</i>	No. of examinees		
				1,000	2,500	5,000
1	0.05	−1.50	0.75	.07 (.03)	.07 (.04)	.07 (.04)
2	0.05	−1.50	1.50	.06 (.04)	.06 (.03)	.08 (.05)
3	0.05	0.00	0.75	.08 (.05)	.10 (.06)	.12 (.08)
4	0.05	0.00	1.50	.17 (.09)	.27 (.19)	.42 (.32)
5	0.05	1.50	0.75	.12 (.07)	.15 (.09)	.22 (.14)
6	0.05	1.50	1.50	.52 (.43)	.83 (.75)	.99 (.97)
7	0.10	−1.50	0.75	.07 (.04)	.09 (.05)	.09 (.05)
8	0.10	−1.50	1.50	.10 (.05)	.11 (.06)	.14 (.08)
9	0.10	0.00	0.75	.11 (.05)	.14 (.08)	.20 (.13)
10	0.10	0.00	1.50	.31 (.21)	.55 (.42)	.78 (.69)
11	0.10	1.50	0.75	.17 (.12)	.29 (.20)	.43 (.33)
12	0.10	1.50	1.50	.73 (.64)	.97 (.94)	1.00 (1.00)
13	0.15	−1.50	0.75	.09 (.05)	.10 (.06)	.11 (.06)
14	0.15	−1.50	1.50	.10 (.06)	.14 (.09)	.18 (.12)
15	0.15	0.00	0.75	.16 (.09)	.17 (.10)	.30 (.20)
16	0.15	0.00	1.50	.40 (.31)	.75 (.64)	.93 (.88)
17	0.15	1.50	0.75	.22 (.13)	.33 (.24)	.58 (.45)
18	0.15	1.50	1.50	.80 (.72)	.99 (.98)	1.00 (1.00)
19	0.20	−1.50	0.75	.09 (.05)	.12 (.07)	.15 (.08)
20	0.20	−1.50	1.50	.15 (.08)	.20 (.12)	.23 (.15)
21	0.20	0.00	0.75	.16 (.09)	.22 (.15)	.36 (.25)
22	0.20	0.00	1.50	.51 (.41)	.84 (.77)	.98 (.97)
23	0.20	1.50	0.75	.25 (.17)	.39 (.28)	.66 (.55)
24	0.20	1.50	1.50	.83 (.76)	.99 (.99)	1.00 (1.00)
25	0.25	−1.50	0.75	.07 (.04)	.14 (.09)	.17 (.10)
26	0.25	−1.50	1.50	.13 (.08)	.21 (.15)	.29 (.21)
27	0.25	0.00	0.75	.16 (.10)	.25 (.17)	.41 (.29)
28	0.25	0.00	1.50	.58 (.47)	.90 (.84)	.99 (.99)
29	0.25	1.50	0.75	.22 (.15)	.47 (.34)	.70 (.58)
30	0.25	1.50	1.50	.84 (.76)	1.00 (.99)	1.00 (1.00)

Note. Proportion of rejections calculated as the number of observed G^2 values exceeding the upper $\alpha = .05$ critical value for a 50:50 mixture of χ^2 distributions with 0 and 1 degrees of freedom out of the 1,000 replications. Values in parentheses represent the proportion of rejections if the standard $\chi^2(1)$ reference distribution were to be incorrectly used. 2PL = two-parameter logistic; 3PL = three-parameter logistic; LR = likelihood ratio.

parameters of 1.50, and differing only in the fact that Item 29 has a discrimination parameter of 0.75, whereas Item 30 has a discrimination parameter of 1.50. With 5,000 examinees, the LR test to detect the presence of a guessing parameter in Item 29 had an estimated power of 70% when the correct reference distribution was used but only 58% power when the incorrect standard reference distribution was used to assess significance. Comparatively, this same test for Item 30 had an estimated 100% power when either reference distribution was used.

Table 2 gives power estimates for the same hypothesis tests as Table 1, but with a Bonferroni correction applied. These results would be of interest when all items of a test of moderate length are being considered for reduction from a 3PL to a 2PL model. From the results in Table 2, it is clear that these tests were only powerful in the detection of a guessing parameter for items that were highly discriminating and at least moderately difficult. For such items, there was again a noticeable difference between the power estimates according to the standard $\chi^2(1)$ reference

Table 2. Proportion of Rejections of the Misspecified 2PL Model in LR Tests Comparing the 2PL and 3PL Models for Each Item of a 30-Item Test with a Bonferroni Correction.

Item	c	b	a	No. of examinees		
				1,000	2,500	5,000
1	0.05	−1.50	0.75	.00 (.00)	.00 (.00)	.00 (.00)
2	0.05	−1.50	1.50	.00 (.00)	.01 (.01)	.01 (.01)
3	0.05	0.00	0.75	.01 (.00)	.01 (.00)	.01 (.00)
4	0.05	0.00	1.50	.01 (.01)	.03 (.02)	.09 (.06)
5	0.05	1.50	0.75	.01 (.00)	.01 (.01)	.02 (.02)
6	0.05	1.50	1.50	.11 (.08)	.39 (.32)	.80 (.75)
7	0.10	−1.50	0.75	.00 (.00)	.01 (.00)	.00 (.00)
8	0.10	−1.50	1.50	.01 (.00)	.00 (.00)	.01 (.01)
9	0.10	0.00	0.75	.01 (.01)	.01 (.00)	.02 (.01)
10	0.10	0.00	1.50	.04 (.03)	.10 (.07)	.33 (.27)
11	0.10	1.50	0.75	.01 (.01)	.03 (.02)	.07 (.04)
12	0.10	1.50	1.50	.29 (.23)	.76 (.70)	.99 (.98)
13	0.15	−1.50	0.75	.01 (.00)	.01 (.00)	.01 (.00)
14	0.15	−1.50	1.50	.01 (.00)	.01 (.01)	.02 (.01)
15	0.15	0.00	0.75	.01 (.01)	.01 (.01)	.03 (.02)
16	0.15	0.00	1.50	.07 (.04)	.28 (.22)	.61 (.52)
17	0.15	1.50	0.75	.02 (.01)	.05 (.04)	.12 (.09)
18	0.15	1.50	1.50	.38 (.32)	.88 (.81)	1.00 (1.00)
19	0.20	−1.50	0.75	.01 (.00)	.00 (.00)	.01 (.00)
20	0.20	−1.50	1.50	.01 (.01)	.02 (.01)	.02 (.01)
21	0.20	0.00	0.75	.01 (.01)	.03 (.02)	.05 (.04)
22	0.20	0.00	1.50	.12 (.09)	.42 (.36)	.82 (.78)
23	0.20	1.50	0.75	.02 (.02)	.07 (.05)	.19 (.14)
24	0.20	1.50	1.50	.44 (.36)	.90 (.86)	1.00 (1.00)
25	0.25	−1.50	0.75	.01 (.00)	.00 (.00)	.01 (.00)
26	0.25	−1.50	1.50	.01 (.01)	.03 (.02)	.05 (.03)
27	0.25	0.00	0.75	.01 (.01)	.03 (.02)	.06 (.04)
28	0.25	0.00	1.50	.15 (.11)	.51 (.44)	.89 (.86)
29	0.25	1.50	0.75	.03 (.02)	.08 (.06)	.21 (.16)
30	0.25	1.50	1.50	.39 (.31)	.89 (.84)	1.00 (1.00)

Note. Proportion of rejections calculated as the number of observed G^2 values exceeding the upper $\alpha = .05/30$ critical value for a 50:50 mixture of χ^2 distributions with 0 and 1 degrees of freedom out of the 1,000 replications. Values in parentheses represent the proportion of rejections if the standard $\chi^2(1)$ reference distribution were to be incorrectly used. 2PL = two-parameter logistic; 3PL = three-parameter logistic; LR = likelihood ratio.

distribution and the correct null distribution, which is a 50:50 mixture of the $\chi^2(0)$ and $\chi^2(1)$ distributions. Thus, these results suggest that LR tests comparing the 2PL and the 3PL models for multiple items of a test will be most powerful in the case of moderate to large samples (e.g., 2,500 examinees or more), highly discriminating (e.g., $a_j \geq 1.50$) and at least moderately difficult (e.g., $b_j \geq 0$) items, and when significance is assessed according to a 50:50 mixture of $\chi^2(0)$ and $\chi^2(1)$ distributions as opposed to the standard reference distribution.

Empirical Example: Comparing the 2PL and 3PL Models for the Items on a Statewide Mathematics Test via an LR Test

As the simulation studies highlighted, the standard χ^2 reference distribution used for LR tests does not hold when comparing the 2PL and 3PL models, as the null hypothesis sets the guessing parameter equal to its lower bound of 0. As such, an adjustment must be made to the standard

LR testing procedure for Type I error rates to be equal to the stated α level and to avoid a potential loss of power. One such method using item response data from a statewide mathematics test will be demonstrated.

Data

The item response data used in this example came from the 2003 *Florida Comprehensive Assessment Test* (FCAT) ninth-grade mathematics test (Florida Department of Education, 2003). This exam was administered to 211,601 students, and in this study the responses from a random sample of 5,000 of these students were analyzed. There were a total of 44 items on this exam, of which 29 were dichotomously scored multiple-choice items. For this example, the authors were concerned with modeling the responses to these 29 multiple-choice items.

Method

To begin the analysis, a 3PL model was fit to all 29 of the multiple-choice items on the 2003 *FCAT Mathematics Test* for Grade 9. The question of interest was then whether or not reduction to a 2PL model would be appropriate for any of these 29 items. To address this question, a technique that is commonly referred to as *backward elimination* in the statistical modeling literature was implemented (e.g., Kutner, Nachtsheim, Neter, & Li, 2005). That is, the initial model included guessing parameters for all 29 items. Then, at each step in the analysis, the LR test statistics corresponding to a test of whether the guessing parameter for each item was significantly different from 0 was calculated. The guessing parameter with the least significant LR statistic was then removed from the model and the process continued until all remaining guessing parameters had statistically significant LR statistics.

The statistical significance of the LR statistics was assessed according to a 50:50 mixture of χ^2 distributions with 0 and 1 degrees of freedom. That is, when the observed LR test statistic G^2 equals t , the p value was given as follows: $1/2[P(X_0 > t)] + 1/2[P(X_1 > t)] = 1/2[P(X_1 > t)]$, where X_0 denotes a random variable with a χ^2 distribution with 0 degrees of freedom and X_1 denotes a random variable with a χ^2 distribution with 1 degree of freedom. The p value is therefore half the p value that would be obtained if the standard $\chi^2(1)$ reference distribution were used. In each step, a Bonferroni correction was used to control the familywise error rate at .05. That is, the results were considered statistically significant if the p value was less than $.05/K$, where K is the number of LR statistics calculated in that step. This approach allowed for the possibility that the most parsimonious model for the test was a 3PL for some items and a 2PL for other items.

Results

The item parameter estimates for the initial model specifying a 3PL for all items are given in Table B1 (see the online appendix). Note that these items generally had larger estimated item discrimination parameters, implying that the LR test should have adequate power to detect the presence of a guessing parameter, based upon the results of the power analysis simulation study. In Table B2, the LR test statistics for the first step in the backward elimination procedure are listed. There were three guessing parameters (c_2 , c_4 , and c_5) whose removal in no way reduced the fit of the model. Thus, in the interest of arriving at the most parsimonious model, one of these parameters, c_2 , was selected for removal from the model. In subsequent steps of the analysis, the parameters c_4 , c_5 , c_1 , c_7 , c_{19} , c_3 , c_{25} , c_{29} , and c_{18} were removed from the model, in that order. The item parameter estimates for the final model are also given in Table B1.

Each time a guessing parameter was removed from the model for the *FCAT* data, the item was subsequently estimated to be more discriminating and less difficult. Such information might potentially affect decisions regarding the use of these items in future administrations of the *FCAT*, illustrating how the choice of a model for a set of item response data can affect conclusions made in subsequent analyses. Furthermore, item parameter estimates will affect reported ability estimates and their standard errors. Figure B1 provides a comparison between the standard error of the ability estimates across the range of θ_i for the initial and final models for the *FCAT* data. The differences are most notable for the lower ability levels, wherein the ability estimates from the final model were more precise than those from the initial model. Such differences underscore the importance of selecting the most appropriate IRT model.

Discussion

The results of this study illustrate that an issue widely recognized in the mixed-effects modeling literature, namely, that when testing if a parameter is equal to one of its boundary values the standard χ^2 reference distribution is no longer valid, also affects the LR tests for IRT model selection presented in many prominent IRT texts. Specifically, when comparing the 2PL and 3PL models for a set of item response data, the LR statistic no longer has a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters estimated in the two models, as the null hypothesis of such a test places the guessing parameters at their lower bound of 0. For testing whether several guessing parameters are all equal to 0, the correct reference distribution is not known. However, when testing whether a single guessing parameter is equal to 0 via an LR test, the correct reference distribution is a 50:50 mixture of χ^2 distributions with 0 and 1 degrees of freedom. Ignoring the boundary issue can result in the selection of an overly simplified model that does not account for the probability of guessing correctly even when it is a contributing factor in the observed item responses. In the conditions that were considered in this simulation study, such tests were found to have adequate power for appropriately retaining a guessing parameter when items were highly discriminating and at least moderately difficult.

Many established testing programs use a particular IRT model as a matter of policy, and as such do not routinely use LR tests to compare the fit of several different IRT models. However, a general awareness of how the boundary issue affects LR tests for IRT model selection is important in the field of IRT modeling, as researchers choosing to apply IRT methods to instruments used in their own research context rely on the information about LR tests for IRT model selection presented in many of the popular introductory IRT texts and in the documentation of many widely used IRT software programs. The potential consequences of ignoring the boundary issue in LR tests comparing the 2PL and 3PL models will be dependent upon the research questions of interest for a particular study. For example, IRT model comparisons can affect conclusions drawn about the comparability of two tests measuring the same construct (e.g., Bachman et al., 1995); results from model comparisons can have implications for the conclusions reached about the quality of items included on a particular scale (e.g., Reise & Waller, 2003); IRT model choice affects item parameter estimates, which, in turn, might be used to assist in questionnaire development (e.g., Edelen & Reeve, 2007); selection of a misspecified IRT model can potentially lead to inaccuracies in the reported measurement errors and test reliability (e.g., Bergan, 2010). Thus, a complete understanding of the statistical properties of the methods used for IRT model selection is essential.

In conclusion, LR tests comparing the 2PL and 3PL models will need to somehow account for the boundary issue to avoid being overly conservative. One potential method for taking this into account when comparing the fit of a 2PL model with the fit of a 3PL model was presented in the analysis of data from the 2003 *FCAT Mathematics Test*. The authors' approach made use

of tests that considered the removal of just one guessing parameter at each step of the analysis, thus allowing for the use of a known reference distribution, and allowing greater modeling flexibility than an approach which assumes that *all* items on a test either need a guessing parameter or do not need one. Alternatively, the ltm package for the statistical computing software R (Rizopoulos, 2006) has the option to request p values from a simulated null distribution for LR tests for IRT model selection, which would be more appropriate than using the standard reference distribution when such tests involve determining whether guessing parameters are significantly different from 0.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

The online appendix is available at <http://apm.sagepub.com/supplemental>

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723.
- Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38, 123-140.
- Bachman, L. F., Davidson, F., Ryan, K., & Choi, I.-C. (1995). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study*. Cambridge, UK: University of Cambridge Local Examinations Syndicate and Cambridge University Press.
- Bergan, J. R. (2010). *Assessing the relative fit of alternative item response theory models to the data*. Retrieved from <http://www.ati-online.com/pdfs/researchK12/AlternativeIRTModels.pdf>
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.
- De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.
- DeMars, C. (2010). *Item response theory*. Oxford, UK: Oxford University Press.
- du Toit, M. (Ed.). (2003). *IRT from SSI*. Lincolnwood, IL: Scientific Software International.
- Edelen, M., & Reeve, B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16, 5-18.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Florida Department of Education. (2003). Florida Comprehensive Assessment Test. Unpublished instrument.
- Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, 64, 273-294.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, 31, 331-358.

- Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). New York, NY: McGraw-Hill.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores* (with contributions by A. Birnbaum). Reading, MA: Addison-Wesley.
- Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G2(dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, 41, 55-64.
- Oehlert, G. W. (2000). *A first course in design and analysis of experiments*. New York, NY: Freeman.
- Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50-64.
- Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, 8, 164-184.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17, 1-25.
- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82, 605-610.
- Silvapulle, M. J., & Sen, P. K. (2005). *Constrained statistical inference: Order, inequality, and shape constraints*. Hoboken, NJ: Wiley.
- Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50, 1171-1177.
- Thissen, D. (1991). *MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory*. Chicago, IL: Scientific Software International.
- Waller, M. I. (1981). A procedure for comparing logistic latent trait models. *Journal of Educational Measurement*, 18, 119-125.
- Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9, 60-62.
- Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245-262.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125-145.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software International.