

Research Paper

Assessing the Relative Fit of Alternative Item Response Theory Models to the Data

by
John Richard Bergan, Ph.D.



**Assessment
Technology
Incorporated**

Assessment Technology Incorporated

6700 E. Speedway Blvd.
Tucson, Arizona 85710

Phone: 520/323-9033 • Fax: 520/323-9139

© 2010 Assessment Technology Incorporated

Assessing the Relative Fit of Alternative Item Response Theory Models to the Data

By John Richard Bergan, Ph.D.
Assessment Technology Incorporated

Table of Contents

Table of Contents.....	i
Acknowledgements.....	ii
I. Item Response Theory and Model Selection	1
II. The Philosophy and Science of Model Selection.....	6
III. References	7

*Copyright © 2010 by Assessment Technology, Incorporated
All rights reserved. No part of this document may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission from the publisher.*

*Assessment Technology, Inc., Publishers
Tucson, Arizona, U.S.A.
Printed in the United States of America.
VI-03162010*

Acknowledgements

I give special thanks to Dave Thissen for his insightful review of this manuscript. I also wish to thank Christine Burnham for her assistance with the analysis of the data. Finally, I want to thank John Robert Bergan, Kathryn Bergan, Christine Burnham, and Sarah Callahan for their many helpful comments regarding the manuscript.

I. Item Response Theory and Model Selection

The standards-based education movement has produced widespread demand for reliable and valid assessments of student ability that can be used to inform instruction. Item Response Theory (IRT) is now widely used to produce scale scores for these assessments. An important aspect of the IRT approach is the selection of an IRT model to represent the data. Model selection is often based solely on philosophical considerations. However, model selection may also be informed by empirical tests. This paper illustrates an empirical approach to model selection. In the process the discussion explores the age-old tension between science and other philosophical perspectives.

Science entails the development of theories involving hypotheses that can be tested through the examination of data. The IRT mathematical model used to produce estimates of item parameters and ability is in fact a testable theory. That theory asserts that observed response patterns for a given set of items can be explained by an unobserved latent variable, which we often call ability or proficiency. Although direct tests of the IRT model become increasingly difficult as the number of items and possible response patterns to those items becomes large, tests of the relative fit of alternative models are possible even with a large number of possible response patterns. In my view, it is important to conduct tests of the relative fit of IRT models involving different item parameters to the data. The reason is that this type of test provides empirical evidence that can inform the selection of an IRT model to represent the data being analyzed. As I will show, tests of relative fit can assist in avoiding seriously misleading interpretations of data that may occur when a misspecified model is selected to represent the data under examination. Yet it appears that tests of relative fit are not widely conducted. The discussion that follows illustrates procedures that can be used to test the relative fit to the data of three IRT models, the one-parameter logistic model (1PL), the two-parameter logistic model (2PL), and the three-parameter logistic model (3PL). In this illustration, these procedures are implemented using the Multilog computer program developed by David Thissen (Thissen, Chen, & Bock, 2003). One of the many advantages to Multilog is that it is designed to accommodate tests of relative fit of the type being examined here.

To illustrate the examination of relative fit, Christine Burnham, an Assessment Technology Incorporated (ATI) Research Scientist, and I used Thissen's Multilog computer program to examine data from a 44 item 5th grade math test administered to 3,098 Arizona students. The analysis related to the 1PL model requires special comment. The one-parameter model was originally developed by Georg Rasch and has been named the *Rasch Model* in his honor. The one-parameter model specified by Rasch makes no assumptions about the shape of the ability distribution or its mean and standard deviation. The model implicitly sets the discrimination parameter to one. In addition, the sum of the difficulty parameters is set to zero. These constraints affect the mean and standard deviation of the ability distribution. Multilog assumes a normal distribution with a mean of zero and a standard deviation of one. We used the Multilog approach to set the ability scale for the 1PL model and the other models that were tested. For the 1PL model (labeled model M1), 44 difficulty parameters and one discrimination parameter were estimated. For the 2PL model (labeled M2), 44 difficulty parameters and 44 discrimination parameters were estimated. Accordingly, the comparison between 1PL model and the 2PL model involved 43 degrees of freedom.

The comparison between the 2PL and the 3PL model (labeled M3) involved an additional 44 parameter estimates for the pseudo guessing parameter.

The three models under examination here reflect a nested hierarchy in which all of the parameters estimated for the 1PL model are included in the parameter set estimated for the 2PL model and all of the parameters estimated for the 2PL model are included in the parameter set for the 3PL model. As a result, difference chi-squares can be computed for each comparison using the likelihood-ratio chi-squared statistic. The significance of these chi-square tests can be determined by referring the results to the chi-square distribution. In general, in science, the most parsimonious model (i.e. the model involving the least number of estimated parameters) is preferred to represent the data. However, additional parameters are justified if they significantly improve the fit of the model to the data. In terms of the three models under discussion here, the 1PL model is the most parsimonious in that the only parameters estimated are the difficulty parameters for each of the items and one discrimination parameter for all of the items. Likewise, the 2PL model which includes both difficulty and discrimination parameters is more parsimonious than the 3PL model, which includes difficulty, discrimination, and pseudo guessing parameters.

The results for the comparisons among the three models are given in Table 1. The table shows the models being compared and the chi-square values, degrees of freedom, and probability levels for each of the comparisons. All of the comparisons yielded results that were significant beyond the .001 level. The conclusion to be drawn from this illustration is that for this assessment the 3PL model is preferred over the 1PL and 2PL models because the 3PL model offers a significant improvement in the fit of the model to the data over the alternative models. In other words, the additional parameters estimated in the 3PL model are justified because they help provide a better fit to the data.

Table 1. Chi-squared tests comparing the fit of the one-, two-, and three-parameter models to the data

Comparison	χ^2	df	p values
M1-M2	1,532.7	43	< .001
M2-M3	833.0	44	< .001
M1-M3	2,365.7	83	< .001

The above results support selection of the 3PL model. The other two models represent cases of model misspecification. It is informative to examine the effect of model misspecification on item parameter estimation. To illustrate the possible effects on item parameter estimates, the comparison between model M1 (the 1PL model) and model M3 (the 3PL model) is considered. Table 2 shows both the 1PL model and the 3PL model item difficulty estimates (b values) for each of the 44 items in the assessment.

Table 2. Item difficulties for the one-parameter model and three-parameter model

Item	1	2	3	4	5	6	7	8	9	10	11
M1 b value	-0.89	-1.04	0.65	-1.73	-0.41	0.83	0.41	-0.14	-0.30	-1.33	-1.52
M3 b value	-0.07	0.13	0.78	-0.41	0.28	0.79	0.55	0.35	0.23	-0.56	-0.60
Difference	0.82	1.17	0.13	1.32	0.69	-0.04	0.14	0.49	0.53	0.77	0.92

Item	12	13	14	15	16	17	18	19	20	21	22
M1 b value	-0.81	-0.66	-0.20	-0.26	-1.28	-0.99	-2.08	-0.31	-1.54	-1.47	-0.45
M3 b value	-0.28	-0.13	0.48	0.27	-0.80	-0.41	-1.49	0.29	-1.00	-1.12	0.44
Difference	0.53	0.53	0.68	0.53	0.48	0.58	0.59	0.60	0.54	0.35	0.89

Item	23	24	25	26	27	28	29	30	31	32	33
M1 b value	0.44	-0.04	-0.48	1.32	0.56	-2.47	1.10	-0.34	-0.12	0.07	0.04
M3 b value	1.42	0.78	-0.17	1.50	1.06	-1.78	1.20	0.39	1.04	0.87	0.83
Difference	0.98	0.82	0.31	0.18	0.50	0.69	0.10	0.73	1.16	0.80	0.79

Item	34	35	36	37	38	39	40	41	42	43	44
M1 b value	-0.29	-2.14	-2.18	-0.58	-0.78	-1.02	-1.67	0.33	-1.93	-1.50	-0.27
M3 b value	0.32	-1.54	-1.42	-0.34	-0.21	-0.56	-1.17	0.76	-1.33	-1.26	0.22
Difference	0.61	0.60	0.76	0.24	0.57	0.46	0.50	0.43	0.60	0.24	0.49

Note the marked differences in difficulty parameter estimates between the two models. The values for the difficulty parameter estimates under the 1PL model are uniformly lower than the estimates computed under the 3PL model. Given the compelling evidence that the 3PL model is preferred over the 1PL model to represent the data, it can be concluded that the values obtained for the 1PL model are highly misleading. Overall the items are more difficult than the 1PL model would lead one to believe. While both the 1PL and 3PL models have parameters that are commonly called “difficulty” parameters, they are not the same: The 1PL model’s “difficulty” parameter reflects the proportion of examinees who respond correctly, while the 3PL’s “difficulty” parameter reflects the proportion of examinees who actually know the answer to the question (setting aside those who gave a correct response by guessing). For multiple-choice items, those two values are not the same. For example, they are especially not the same for item 2 (in Table 3 on page 4) that has a model-estimated 0.45 probability of a guessed correct response for examinees who do not know the answer.

One of the most important considerations in item evaluation involves measurement error. Figure 1 shows the standard error of measurement curves for the 1PL model and the 3PL model across the ability continuum. The figure also shows the Test Information curves for both models across the ability continuum. In addition, the figure shows the marginal reliability coefficients for each model.

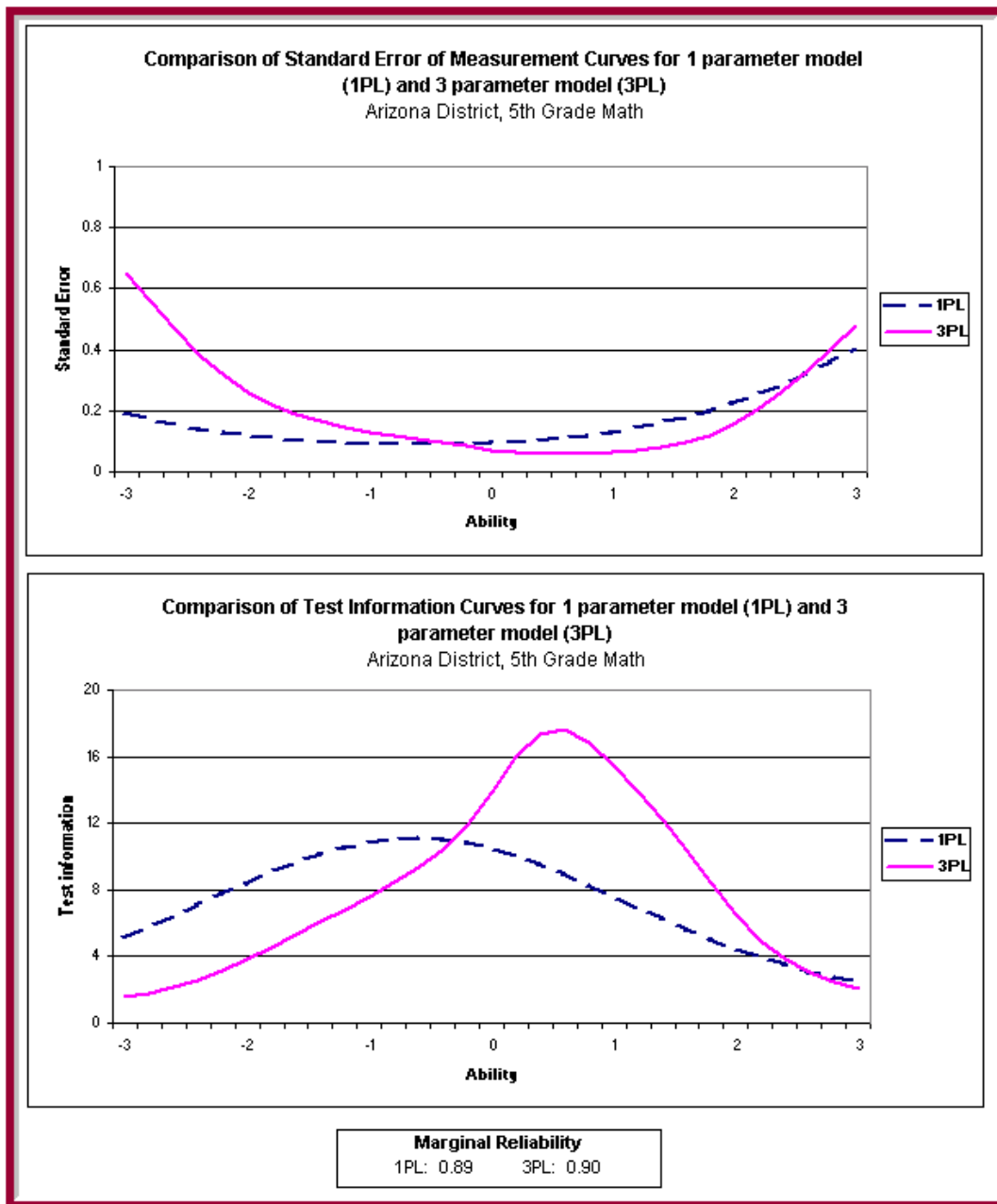


Figure 1. Comparison of standard error and test information for the one-parameter and three-parameter models

There is virtually no difference in the reliability coefficients for the two models. However, the test information curves and measurement error curves differ markedly. The test information curve for the 3PL model indicates that test information is maximal for students whose ability levels are above average. The curve for the 1PL model shows maximum test information for ability levels that are below average. This is very misleading.

In addition to representing the data better, the 3PL model provides more useful information about the items than the 1PL model. In particular, it provides not only a difficulty parameter, but also a discrimination and a guessing parameter for each item. Table 3 shows the item discrimination, difficulty, and pseudo guessing parameter estimates for the 3PL model for each of the items in the test described above.

Table 3. Item parameter estimates for the 3PL model.

Item	Discrimination	Difficulty	Guessing	Item	Discrimination	Difficulty	Guessing
1	1.18	-0.07	0.32	23	0.76	1.42	0.26
2	0.80	0.13	0.45	24	0.79	0.78	0.27
3	1.58	0.78	0.14	25	0.51	-0.17	0.14
4	1.39	-0.41	0.50	26	1.94	1.50	0.17
5	1.58	0.28	0.29	27	1.09	1.06	0.21
6	1.76	0.79	0.10	28	0.89	-1.78	0.20
7	1.64	0.55	0.13	29	1.48	1.20	0.14
8	1.87	0.35	0.23	30	0.65	0.39	0.26
9	2.00	0.23	0.23	31	1.10	1.04	0.38
10	1.13	-0.56	0.27	32	1.03	0.87	0.29
11	1.33	-0.60	0.32	33	0.74	0.83	0.26
12	0.58	-0.28	0.24	34	0.77	0.32	0.23
13	1.08	-0.13	0.20	35	0.93	-1.54	0.14
14	0.67	0.48	0.24	36	1.12	-1.42	0.15
15	0.76	0.27	0.21	37	0.47	-0.34	0.13
16	0.87	-0.80	0.17	38	1.02	-0.21	0.22
17	0.82	-0.41	0.23	39	0.85	-0.56	0.17
18	0.85	-1.49	0.20	40	0.89	-1.17	0.14
19	0.84	0.29	0.24	41	0.93	0.76	0.18
20	0.86	-1.00	0.19	42	0.94	-1.33	0.16
21	0.61	-1.12	0.21	43	0.64	-1.26	0.14
22	0.90	0.44	0.33	44	1.08	0.22	0.20

Note the differences among the discrimination parameter estimates and pseudo guessing parameter estimates obtained for the 3PL model. It is observed that the levels of guessing for some items are higher than desirable. It is also observed that some discrimination parameters are lower than what might be desired. This information may be useful in guiding item selection for use on other assessments.

II. The Philosophy and Science of Model Selection

The results favoring the 2PL model over the 1PL model, and the 3PL model over the 2PL model are not surprising. Consider the comparison between the 1PL model and the 2PL model. The general IRT model can be viewed as a factor analysis model. The discrimination parameter for each item in the 2PL model is functionally related to its factor loading for a unidimensional factor analysis model. The 1PL model assumes that each factor loading is the same as every other factor loading. That is likely to be a rare occurrence. Likewise, consider the comparison between the 2PL model and the 3PL model. The fact that the 3PL model is preferred to represent the data provides compelling evidence that the children responding to the multiple-choice items on this assessment sometimes guess. This is hardly surprising.

Given the empirical support for the 3PL model and the obvious reasons for assuming that the 3PL model would be preferred over the 1PL model, one may ask why there is a significant level of support for the 1PL model in the absence of initiatives designed to compare the fit of the two models to the data. The answer to this question lies in the age-old tension that exists between philosophy and that brand of philosophy that is called science. Thissen and Orlando (2001) describe the Rasch approach to model selection as follows:

Optimal measurement is defined mathematically, and then the class of item response models that yields such measurement is derived. The item-response model is then used as a Procrustean bed that the item-response data must fit, or the item is discarded. Item analysis in this approach consists primarily of analysis of residuals from the model.

Science proceeds in exactly the opposite fashion to the Rasch approach to model selection. In the Rasch approach, data that do not fit the theory expressed in the mathematical model are ignored or discarded. In the scientific approach, theory is discarded or modified if it is not supported by data. Adherence to a scientific approach does not imply that there are no bad items. Indeed, measurement conducted in accordance with the traditional scientific approach facilitates effective item evaluation and selection. At the same time, it calls for an evidence-based approach to model selection.

ATI has named its flagship software application *Galileo K-12 Online* in recognition of the fact that the tension between science and philosophy continues to exist in education and educational measurement. Despite this tension, science generally prevails in the end. However, as the story of Galileo reminds us, prevailing may sometimes take nearly 400 years. In my view, science prevails not only because it increases our understanding of life, the physical world, and the universe beyond, but also because it is expensive and risky to ignore objective data. It is important for us to understand that there really are rings around Saturn, that the Earth really does orbit the Sun, and that sometimes children really do guess on multiple-choice tests.

III. References

Thissen, D., Chen, W-H, & Bock, R.D. (2003). *Multilog (version 7)*. Lincolnwood, IL: Scientific Software International.

Thissen, D., & Orlando, M. (2001) *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.