Comparisons of Item Response Theory Algorithms on Force Concept Inventory

Li Chen[1, 2], Jing Han[1], Jing Wang[3], Yan Tu[2*] and Lei Bao[1*]

[1] *The Ohio State University, Columbus, OH, USA*

[2] *Southeast University, Nanjing, Jiangsu, China*

[3] *Eastern Kentucky University, Richmond, KY, USA*

(Received 23 August 2011; accepted 19 December 2011)

Item Response Theory (IRT) is a popular assessment method widely used in educational measurements. There are several software packages commonly used to do IRT analysis. In the field of physics education, using IRT to analyze concept tests is gaining popularity. It is then useful to understand whether, or the extent to which, software packages may perform differently on physics concept tests. In this study, we compare the results of the 3-parameter IRT model in R and MULTILOG using data from college students on a physics concept test, the Force Concept Inventory. The results suggest that, while both methods generally produce consistent outcomes on the estimated item parameters, some systematic variations can be observed. For example, both methods produce a nearly identical estimation of item difficulty, whereas the discrimination estimated with R is systematically higher than that estimated with MULTILOG. The guessing parameters, which depend on whether "pre-processing" is implemented in MULTILOG, also vary observably. The variability of the estimations raises concerns about the validity of IRT methods for evaluating students' scaled abilities. Therefore, further analysis has been conducted to determine the range of differences between the two models regarding student abilities estimated with each. A comparison of the goodness of fit using various estimations is also discussed. It appears that R produces better fits at low proficiency levels, but falls behind at the high end of the ability spectrum.

I. INTRODUCTION

Item Response Theory (IRT) is a family of models, based on a statistical framework, which provide stable estimates of item and examinee parameters (Junker, 1999; Yen, 2006). Recently, IRT has gained attention in physics and astronomy education (Lee et al, 2006; Wang and Bao, 2010; Wallace and Bailey, 2010). IRT is capable of providing scaled latent score estimates for individual examinees, which enhances the information obtained through testing and helps researchers and educators evaluate teaching and learning effectiveness.

One widely used IRT model is the three-parameter logistic model (3PL model), in which the possible measurement outcome of an item is described with a probability function of the item characteristics and students' proficiency:

$$P(\theta) = c + \frac{1-c}{1+\exp[-1.7a(\theta-b)]} \qquad (1).$$

Here, $P(\theta)$ is the probability for a student with ability θ to correctly answer a question. The assessment characteristics of a question are described in terms of three parameters, the item discrimination a, the item difficulty b, and the guess parameter c. The parameters are usually obtained with a large scale data set through a regression estimation process using Marginal Maximum Likelihood (MML) algorithms (Matthew S. Johnson, 2007, Dimitris Rizopoulos, 2006).

Since the estimation processes are quite complicated, the computation is usually done with existing commercial and open-source software packages for IRT analysis. The most popular ones include R (with its LTM package), MULTI-LOG, PARSCALE, BILOG, ASCAL, LOGIST. Usually, the specific algorithms used by a software package are not publicly available. Due to different computational approaches the nature of the complex numerical manipulations involved, different software packages rarely yield identical outcomes. It is important to understand in what ways these results may differ and how such differences may impact our interpretations of the assessment results.

Many studies have compared software packages different situations, with mixed outcomes. Demars (2001) studied PARSCALE and MULTILOG and showed that the two packages have consistent performance in estimating item parameters in a number of simulated situations. Jurich's research (2009) suggested that the performance of a freeware, Hanson's IRT Command Language (ICL), is equally as effective as PARSCALE on parameter estimation under all conditions. Comparisons of ASCAL and LOGIST in item parameter estimating (Gary Skaggs, 1989), however, suggested that the difference between the two packages depends on conditions such as sample size and the number of questions in a test.

In this paper, we compare MULTILOG and R. MULTILOG is commercial software for IRT analysis. There are two options for using MULTILOG, one is a straight run of IRT regression and the other involves pre-processing to impose Gaussian prior distribution for item parameters before the IRT regression starts (Mathilda Du Toit, 2003). Therefore, it is important to understand how the pre-processing may influence the results.

R is an open-source free programming language for statistics. It contains a free package, LTM, which can be used to do IRT analysis through Latent Trait Model (Dimitris

© iSTARClass Ltd.

Rizopoulos, 2009). There is no option to do pre-processing in R. Thus, in this study, we compare the performance of three specific algorithms, R, MULTILOG without pre-processing (M_P), and MULTILOG with pre-processing (M_{NP}), on a popular physics concept test.

In physics education, standardized conceptual tests have been widely used in research and education practices. The Force Concept Inventory (FCI) is the most commonly used instrument and has led to many groundbreaking studies (Hestenes, et al. 1992; Hake, 1998). The FCI contains 30 questions in multiple choice format (5 choices each question) covering approximately a dozen of core concepts in introductory mechanics. Research has shown that the FCI is appropriate for IRT analysis (Wang and Bao, 2010), which can provide better insight for interpreting the assessment results than classical test theory.

Because the IRT method is gaining popularity in physics education, it is important to know how different algorithms may affect the analysis outcomes. In responding to this research question, this study builds on the existing work to compare R and MULTILOG on their performances in analyzing FCI data.

II. METHOD

At the Ohio State University, from September 2003 to June 2007, the students who enrolled in calculus-based introductory mechanics courses took a pre-FCI test in the second week and a post-FCI test in the week before final exams. This data collection effort was a part of regular lab activities until September 2007, when a reformed lab curriculum was implemented. The average pre- and post-FCI scores for each quarter (about 200 to 300 students per quarter) remained fairly steady over time. In this study, the analysis is done with the pre-test data only, which contains 3139 data points. We combined all the pre-test data over the years to conduct the analysis. Students' responses to each item are coded into a 1-0 binary form for correct and incorrect answers. The average score of this population is 49.27% with a standard deviation of 18.13%. To check the normality of the score distribution, a "Quantile-Quantile" plot was shown in Figure 1. The results suggest that student scores follow reasonably a normal distribution. The condition of this data set is appropriate for conducting IRT analysis.

To apply the 3PL IRT model, there are two fundamental assumptions about features of the test questions, namely, the unidimensionality and local item independence. The unidimensionality describes whether a test is intended to measure the proficiency level of a common ability. Local item independence assumes that for each examinee, his/her performance on one item is independent of his/her performance on another item. It has been shown that if the unidimensionality assumption is satisfied, the local item independence assumption is automatically verified. (Hambleton & Swaminathan, 1985. Lord & Novick, 1968).

The dimensionality of the FCI test result is examined using eigenvalue analysis of the correlation matrix calculated based on the tetrachoric correlations among 30 test items (Reckase, 1979). The case of a unidimensional correlation matrix should have one eigenvalue much larger than the rest. (Reckase, 1979). The eigenvalues of the FCI pre-text correlation matrix are plotted in Figure 2. The first ei-genvalue is significantly larger than the rest, which suggests a single proficiency accounting for a significant portion of the variances and the assumption of unidimensionality is reasonable.

After confirming the validity of applying the IRT model to the analysis of the FCI results, we implemented the three IRT methods on the same FCI data. In each situation, both student ability and the item parameters were estimated. The results are compared to determine the differences among the three methods and the possible causes of such differences.

III. RESULTS AND ANALYSIS

A. Comparisons of the Item Parameters

As discussed earlier, in the 3PL model, three parameters describe the characteristics of each test items. They are item difficulty (b), item discrimination (a), and guessing (c). It is then important to know whether and how the item parameters produced by different software packages would vary.

Using the pre-test data (N=3139), item parameters for each of the FCI questions are obtained using three methods: R, MULTILOG with pre-processing (M_P), and MULTILOG without pre-processing (M_{NP}). The results are summarized in Table 1 and plotted in Figure 3 for easy comparisons.

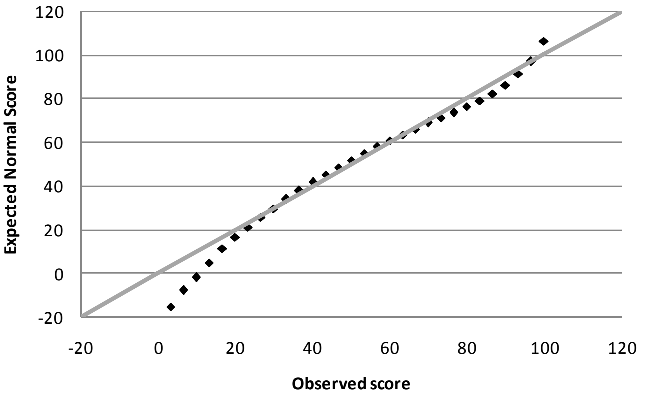


Figure 1. Quantile-Quantile Plot of student pre-test scores on the FCI test.

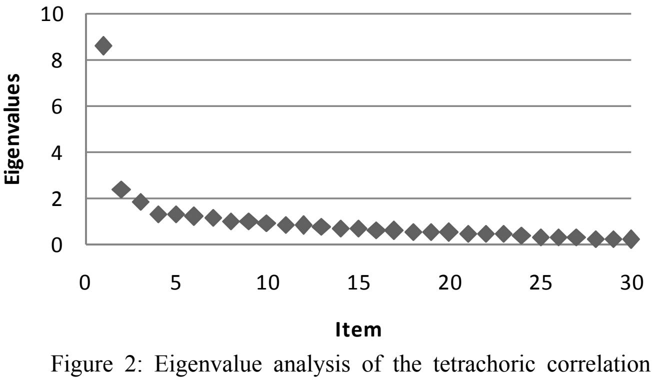


Figure 2: Eigenvalue analysis of the tetrachoric correlation matrix of FCI data.

In general, the item parameters obtained by the three methods are quite similar. In particular, the item difficulty is nearly identical for all three methods (p=0.922). This result agrees with an intuitive response, since the item difficulty determines the center of the fit, which should closely match the mean value of the performance measure (the average score) that remains constant regardless of the regression methods used. A consistent result for item difficulty also indicates that the characteristic curves of an item obtained with different methods are centered together.

For the item discrimination, the results from M_P and M_{NP} are similar, indicating that pre-processing does not affect the estimation processes. However, there are significant differences between R and either M_P or M_{NP} (p=0.004); the discriminations obtained from R are systematically higher than those of M_P and M_{NP} for all 30 items. This result suggests that the slope at the center of the item characteristic curve obtained from R will be steeper than that of the MULTILOG curves.

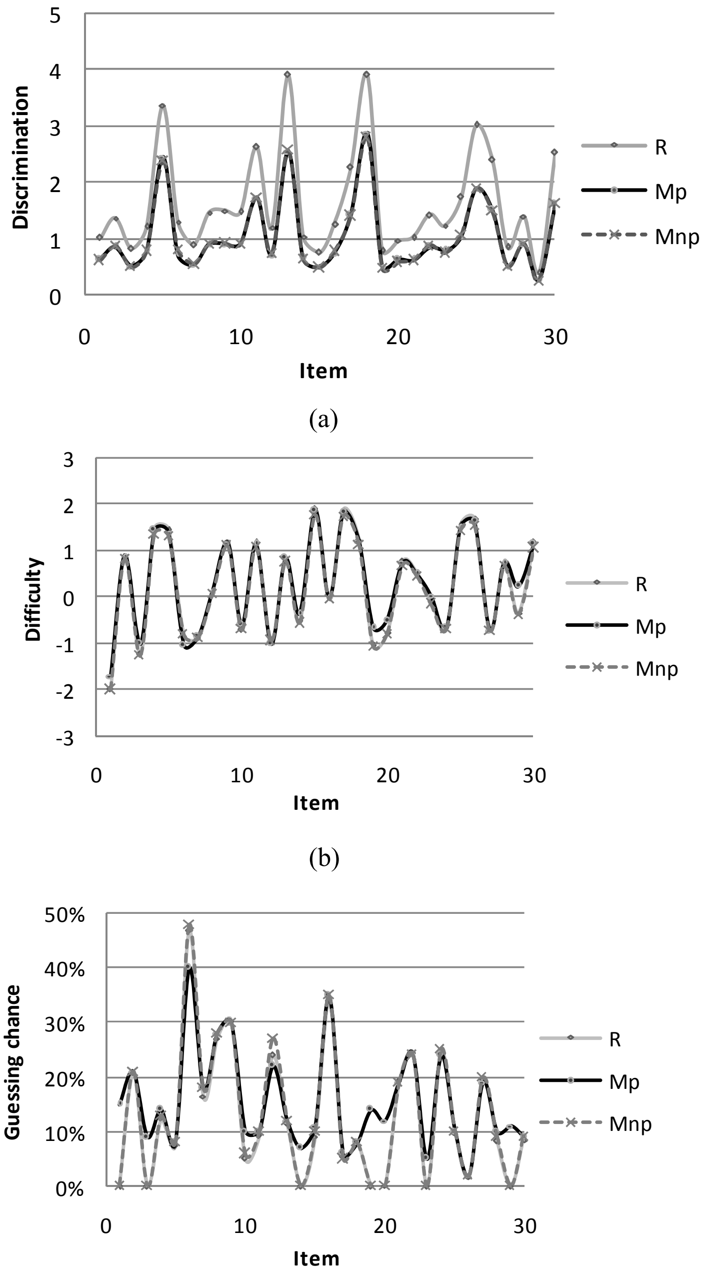
For the guessing parameter, results from R and M_{NP} are nearly identical, while on some items M_P would produce quite different outcomes. The differences occurred on 8 items, for which R and M_{NP} would produce near zero guessing parameters. In contrast, MULTILOG with pre-processing (M_P) consistently produces larger guessing parameters. This result seems to be an outcome of pre-processing, which imposes Gaussian prior distribution for all item parameters, including the guessing chances. Since each FCI question has 5 choices, the center of the Gaussian distribution of the guessing parameter is usually set to be 0.2 (Mathilda Du Toit, 2003). It appears that the pre-processing may have artificially moved up the guessing parameter by favoring a non-zero value.

From research in physics education, it has been widely recognized that students entering introductory college mechanics courses usually have well-established, naïve conceptions about physics. As a result, when students answer incorrectly, they consistently choose the answers that reflect their naïve concepts, rather than guessing (Bao & Redish, 2006; Bao, Hogg, & Zollman, 2002). Therefore, for many FCI ques-tions, close to zero guessing parameters should be allowed. For this reason, we consider pre-processing an unsuitable procedure to use when analyzing concept test items, about which students may have well-developed prior conceptions.

B. The Fit of Item Characteristic Curves

Due to the variations in the item parameters obtained with the three IRT methods, the item characteristic curves will also be different, which in turn will impact the quality of the fit between a model and the data. For a visual comparison of the quality of the fit, the item characteristic curves from the different models for three selected items are plotted in Figure 4.

The three items (questions 1, 5, and 23) are chosen for their different representative features. Question 1 is an easy question with low item discrimination and relatively high guessing chance. Question 5 is a hard question with high discrimination. Question 23 is an average question with a medium level of difficulty and discrimination. The guessing parameters for both question 5 and 23 are quite small.



(a)

(b)

(c)

Figure 3. Comparisons of item parameters estimated with the three methods. The item discriminations, item difficulties, and guessing chances for different items are plotted in (a), (b) and (c) respectively. In all graphs, the x-axis is the item number. The gray line represents the data obtained through R. The black line gives the data using MULTILOG with pre-processing (M_P), and the gray dash line shows the results from MULTILOG without pre-processing (M_{NP}).

Table 1. Item parameters estimates. 30-Items FCI Test

Questions	Discrimination (a)			Difficulty (b)			Guessing (c)		
	R	M_P	M_{NP}	R	M_P	M_{NP}	R	M_P	M_{NP}
Q1	1.02	0.63	0.61	-2.00	-1.71	-2.00	0%	15%	0%
Q2	1.36	0.84	0.86	0.90	0.90	0.83	21%	21%	21%
Q3	0.81	0.50	0.49	-1.24	-1.01	-1.26	0%	9%	0%
Q4	1.21	0.79	0.78	1.48	1.44	1.36	12%	14%	13%
Q5	3.38	2.45	2.37	1.46	1.40	1.32	7%	8%	8%
Q6	1.30	0.72	0.80	-0.84	-1.05	-0.82	47%	40%	48%
Q7	0.89	0.54	0.55	-0.91	-0.86	-0.89	16%	18%	18%
Q8	1.46	0.89	0.90	0.09	0.10	0.05	27%	28%	28%
Q9	1.49	0.90	0.91	1.20	1.17	1.10	30%	30%	30%
Q10	1.47	0.92	0.9	-0.70	-0.59	-0.70	5%	10%	6%
Q11	2.62	1.72	1.72	1.18	1.15	1.07	9%	10%	10%
Q12	1.19	0.70	0.74	-0.98	-1.02	-0.94	24%	22%	27%
Q13	3.91	2.55	2.57	0.84	0.84	0.76	12%	12%	12%
Q14	1.03	0.65	0.63	-0.53	-0.36	-0.56	0%	7%	0%
Q15	0.74	0.50	0.48	1.90	1.87	1.77	9%	11%	10%
Q16	1.25	0.74	0.78	-0.02	-0.03	-0.04	34%	35%	35%
Q17	2.26	1.42	1.41	1.89	1.82	1.74	5%	6%	5%
Q18	3.92	2.88	2.79	1.24	1.21	1.13	8%	8%	8%
Q19	0.78	0.51	0.47	-1.05	-0.66	-1.07	0%	14%	0%
Q20	0.96	0.63	0.58	-0.79	-0.51	-0.81	0%	12%	0%
Q21	1.03	0.62	0.62	0.76	0.75	0.68	19%	19%	19%
Q22	1.42	0.84	0.87	0.52	0.51	0.45	24%	24%	24%
Q23	1.23	0.79	0.75	-0.09	0.03	-0.14	0%	5%	0%
Q24	1.73	1.03	1.06	-0.68	-0.68	-0.69	25%	24%	25%
Q25	3.02	1.89	1.88	1.55	1.50	1.43	10%	10%	10%
Q26	2.40	1.51	1.48	1.68	1.63	1.55	2%	2%	2%
Q27	0.84	0.50	0.51	-0.71	-0.73	-0.74	19%	19%	20%
Q28	1.40	0.89	0.89	0.72	0.74	0.66	8%	10%	9%
Q29	0.39	0.26	0.24	-0.35	0.23	-0.39	0%	11%	0%
Q30	2.53	1.61	1.63	1.18	1.15	1.07	9%	9%	9%

Note: M_P stands for MULTILOG with pre-processing. M_{NP} stands for MULTILOG without pre-processing

Figure 4 contains a 3x3 matrix of item characteristic curves (ICC) overlaid with actual student data. Each row shows three ICC's on a single question obtained from R, M_P, and M_{NP} (from left to right). The student data is calculated by first sorting students from largest to smallest, based on their estimated θ as estimated with the corresponding method. The students are then sequentially put in 31 groups; the first 30 groups contain 100 students each, and the last group contains 139 students. For each group, the mean of the estimated θ and the actual scores are calculated and plotted.

In general, by observing the plotted data points and fitting curves, we can see that all three methods fit the data well, but with slight differences. It seems that MULTILOG gives more weight to the center part of the data and fits the central portion better than R does. On the other hand, R stretches more to fit the two ends, which, as a result, increases the discrimination. Between the two MULTILOG methods (M_P and M_{NP}), the differences in their fits are very small.

C. Reliability of Student Ability Estimations

A common goal of using IRT is to produce an ability scale for all students so that they can be evaluated and compared across different tests and population backgrounds. Therefore, it is of great importance that the different software packages produce consistent outcomes of ability estimates for individual students.

In this study, student ability is estimated together with the item parameters. With each of the three methods, there are 3,139 student ability variables (θ's) and 90 item parameters (a, b, c for each of the 30 questions) to be estimated. Since there are differences among the item parameters obtained from different methods, it can be expected that the estimated student abilities will also vary.

For each of the 3,139 students, three estimated abilities are obtained using the three methods. We then compare the differences among the mean values of the estimated abilities. The results are summarized in Table 2.

We can see that the average θ obtained through R is nearly zero, which indicates that the algorithm used in R anchors the student ability at the center of student performance. On the other hand, MULTILOG methods produce slightly higher estimates of student ability than R does (about 7% of SD for M_P and 26% of SD for M_{NP}). This result is consistent with the item parameter estimations. For a given target probability (score) on an item, student estimated ability is connected with item parameters through Eq. (2), which is rearranged from Eq. (1).

$$\theta = \frac{1}{1.7a}\ln(\frac{P-c}{1-P}) + b \qquad (2).$$

In the previous section, we see that R produces a larger a, similar b, and smaller c compared with M_P, which leads to an overall outcome of larger ability estimated by M_P. Meanwhile, M_{NP} produces a smaller a, similar b, and similar c compared with R, which makes the estimated ability by M_{NP} the largest among the three methods.

To understand the impact of differences between estimated abilities, it is also important to know how such differences may vary across different ability levels. For this analysis, we compare Δθ between M_{NP} and R and M_{NP} and M_P. Here, M_{NP} is chosen as the common base for comparison, as it has only one algorithmic change from either R or M_P. Figure 5 shows the Δθ's plotted against the θ estimated by M_{NP}. Again, each data point gives the mean of a group of 100 students.

From Figure 5, it is clear that θ obtained through M_{NP} is systematically higher than with M_P and R, when θ is smaller than 1.3 (approximately), a value that contains about 94% of the students. The result is inversed when θ is higher than 1.3.

Based on the calculation, we also found a linear relationship that can cross relate the θ obtained with different methods:

$$\begin{aligned} &\theta_{M_P} = 1.038 \times \theta_{M_{NP}} - 0.097 && R^2 = 0.994 \\ &\theta_R = 1.074 \times \theta_{M_{NP}} - 0.164 && R^2 = 0.988 \qquad (3). \\ &\theta_R = 1.034 \times \theta_{M_P} - 0.063 && R^2 = 0.994 \end{aligned}$$

The differences in estimated ability will cause variations in the predicted probabilities of giving correct answers, which we will refer to as predicted scores. Therefore, it is valuable to develop a sense of how differences in estimated ability are reflected in the scale of student scores, which provides useful insights for analyzing both measurement uncertainties and the confidence level of assessment outcomes. For this analysis, we compare the differences between predicted scores obtained with the three IRT methods (see Eq. (4)).

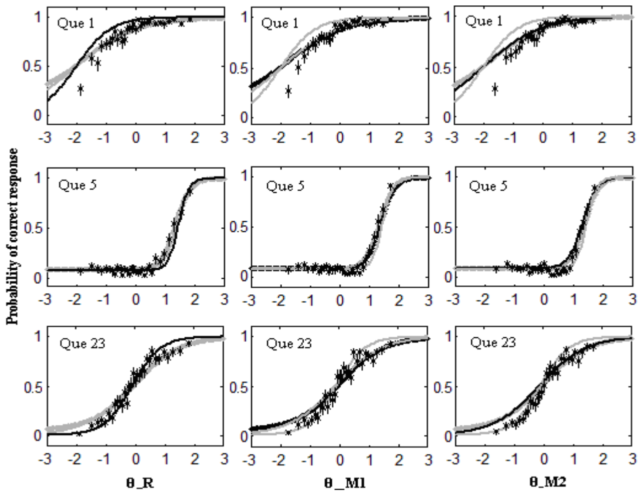


Figure 4. The item characteristic curves (ICC) for three FCI questions obtained through R, MULTILOG with pre-processing (M_P) and MULTILOG without pre-processing (M_{NP}). The y-axis is the probability of giving a correct response. The x-axis is the proficiency θ estimated through the different methods. Each column shows the fits of the three questions with a particular model shown in dark black lines. The gray lines are fits of the other two methods. The data points in each graph show 31 groups of the observed probabilities (student scores) vs. estimated θ that each computed with 100 students' mean value of the estimated θ and the observed score (the last group contains 139 students). The groups are formed with students ordered according to their estimated θ. The error bars of the data points are standard errors for the mean.

Table 2. Student ability estimates using different methods. The means and standard deviations are calculated based on the estimated θ of 3139 students. The standard deviation of the difference is calculated based on 3139 paired differences.

IRT Methods	Mean θ (SD)	Differences	Mean Differences (SD)
R	0.004 (0.922)	M_{NP} – R	0.153 (0.119)
M_P	0.065 (0.888)	M_P – R	0.062 (0.075)
M_{NP}	0.157 (0.853)	M_{NP} – M_P	0.092 (0.076)

$$\Delta P_1 = P_R - P_{M_{NP}} = (1-c)\times\left(\frac{1}{1+\exp\left(-1.7a\left(\theta_{M_{NP}}+\Delta\theta_1-b\right)\right)}-\frac{1}{1+\exp\left(-1.7a\left(\theta_{M_{NP}}-b\right)\right)}\right)$$

$$\Delta P_2 = P_{M_P} - P_{M_{NP}} = (1-c)\times\left(\frac{1}{1+\exp\left(-1.7a\left(\theta_{M_{NP}}+\Delta\theta_2-b\right)\right)}-\frac{1}{1+\exp\left(-1.7a\left(\theta_{M_{NP}}-b\right)\right)}\right) \quad (4).$$

Using Eq. (4), the differences in the predicted scores of 30 FCI questions (total percentage score) are calculated with different IRT methods. The mean values of the results are summarized in Table 3. Figure 6 shows how the differences in the predicted scores may vary with estimated ability and different questions. The results show that the variations in abilities estimated with different methods will cause uncertainties equivalent to about 2 to 3% of the students' scores.

From Figure 6a, we can see that the differences in predicted scores are small at the extremes (low and high θ) and peak when θ equals approximately 0.6. From Figure 6b, we can see that differences in the predicted scores do not vary significantly across different items. In general, the differences are in the range of 1 to 2% for most items, showing a somewhat uniform distribution of differences in predicted scores among the FCI items.

D. Goodness of Fit

In addition to the differences in parameter estimations by the three IRT methods, it is also important to determine how well the individual methods fit the data of the physics concept test. Here, we use the mean error (ME) and root mean square deviation (RMSD) to evaluate the goodness of fit.

To calculate the ME and RMSD, we first arrange the students' data according to their estimated ability, in descending order. Then, the students are divided into 62 proficiency groups, each containing 50 examinees (with the last group containing 89 students). For each item, we can calculate the average observed score and three predicted scores. For each IRT method, the ME and RMSD can be calculated using Eq. (5):

$$ME = \frac{\sum_{k=1}^{62}\sum_{i=1}^{30}\left(P_{ki}-S_{ki}\right)}{N}$$

$$RMSD = \sqrt{\frac{\sum_{k=1}^{62}\sum_{i=1}^{30}\left(P_{ki}-S_{ki}\right)^2}{N-1}} \quad (5).$$

Here S_{ki} and P_{ki} are the average observed and predicted score of the kth group on ith item, respectively.

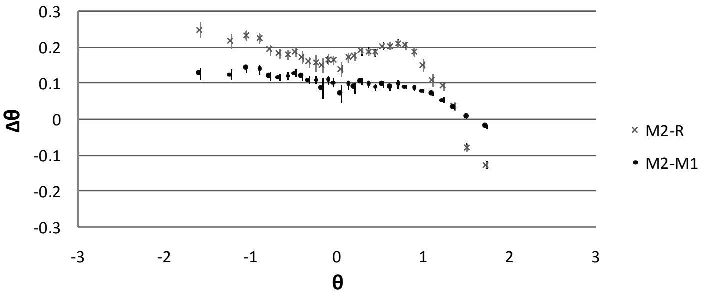
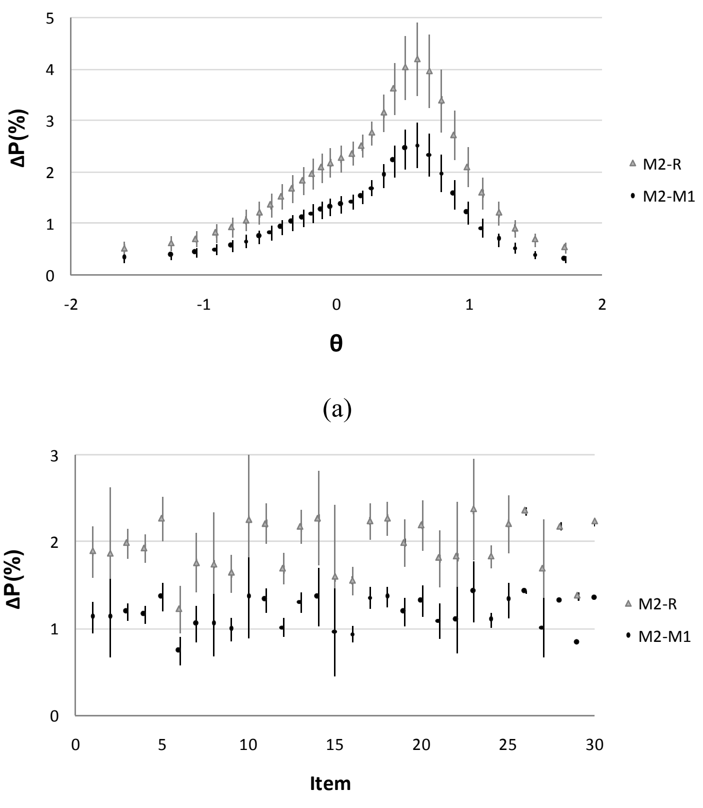


Figure 5. Differences (Δθ's) between different methods. There are 31 groups, each containing 100 data points with the last group containing 139 students. The data points are means of Δθ plotted against means of θ estimated with M_{NP}. The error bars are the standard errors of the mean.

Table 3. The mean differences of the predicted probabilities of answering the FCI test correctly under different IRT methods.

IRT model	Δθ		ΔP
M_{NP} – R	$\Delta\theta_1$–SD_1	0.034	0.43%
	$\Delta\theta_1$	0.153	1.96%
	$\Delta\theta_1$+SD_1	0.272	3.47%
M_{NP} – M_P	$\Delta\theta_2$–SD_2	0.016	0.20%
	$\Delta\theta_2$	0.092	1.18%
	$\Delta\theta_2$+SD_2	0.168	2.15%



(a)

(b)

Figure 6. Differences of the predicted scores using different methods (a) at different θ levels; (b) across different test items. The error bars are standard errors of the means.

The *ME* and *RMSD* for the different IRT methods are summarized in Table 4. We can see that the mean error of R is closer to zero than that of the other two methods, which is consistent with the fact that R's average estimated ability is nearly zero. On the other hand, the two MULTILOG methods both have positive *ME*s, suggesting that the predicted scores are consistently larger the observed scores. With the *RMSD* measure, R produces a larger value than the MULTILOG methods, indicating a wider span of estimated student ability.

For students at different ability levels, the goodness of fit also varies, which is shown in Figure 7a. The results show that towards the lower end of θ, the *ME* of R is very close to zero, smaller than the *ME*'s of the two MULTILOG methods. For θ larger than zero, R and M_P have similar negative *ME*'s, while the *ME* of M_{NP} is mostly above zero.

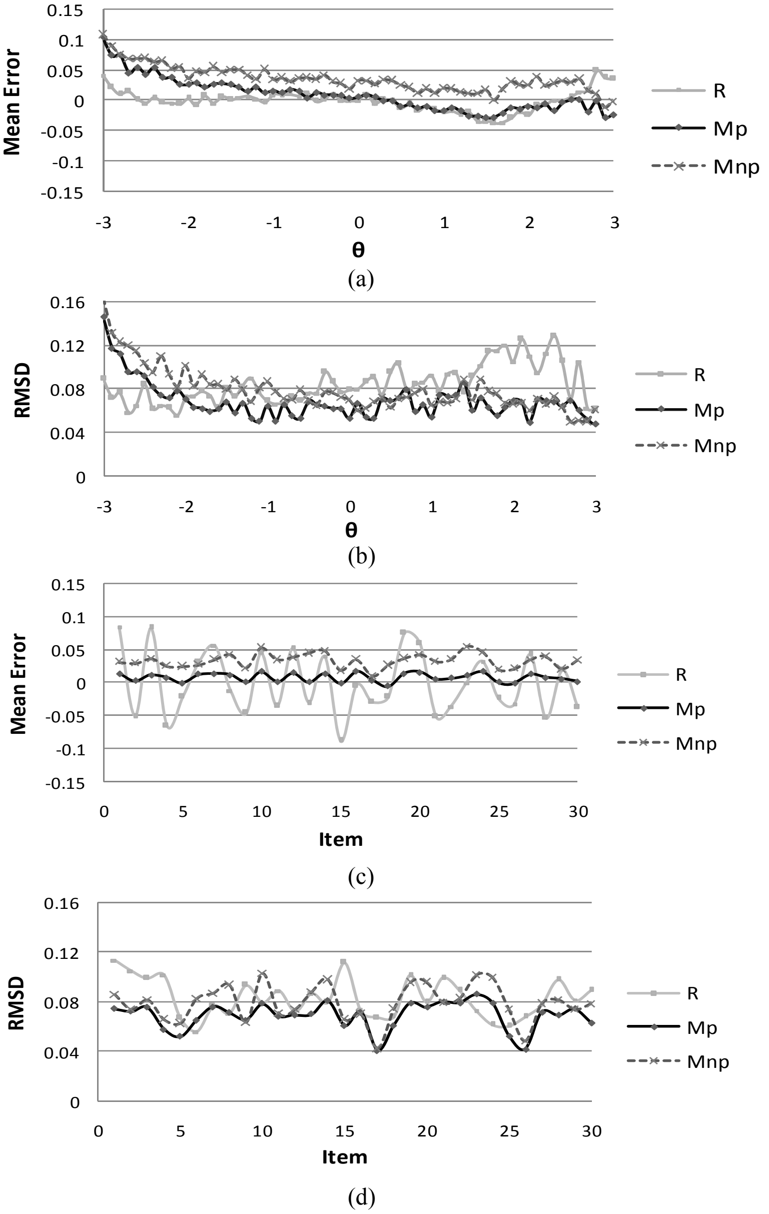
Since the *ME* counts both positive and negative values, it gives a better measure of the center of the fit, but cannot reflect the range of the uncertainties. With the *RMSD* measure, one can see more of the range of the variances between the observed data and the fit. Based on the results in Figure 7a, we can see that the two MULTILOG methods fit well for θ larger than -1.5, while the R method fits well at the lower end of θ but produces a larger RMSD than the MULTILOG methods for θ larger than zero.

Overall, the results show that R produces the fit that best matches the center of the data, while Mp seems to produce the best fit in terms of a smaller range of variances.

Figure 7b shows the results of *ME* and *RMSD* of fit for individual items. The results show that the MULTILOG methods produce more stable mean errors across different items. It appears that MULTILOG with pre-processing produces the best fit based on both the *ME* and *RMSD* measures.

Table 4. Average mean errors and root mean square deviations.

IRT model	Mean error	RMSD
R	-0.001	0.084
M_P	0.008	0.069
M_{NP}	0.033	0.073



(a)

(b)

(c)

(d)

Figure 7. The mean errors and root mean square deviations (a) at different θ levels; (b) across different items.

In summary, R produces a better fit that matches the center of the data with the estimated ability and predicted scores; however, it creates larger range of variances than the two MULTILOG methods. The pre-processing in one of the MULTILOG methods seems to improve the goodness of fit regarding the consistency of the uncertainties across different items, but at the expense of slightly higher predicted scores (at 1% level).

III. SUMMARY AND CONCLUSIONS

In this paper, three 3PL IRT methods, R, MULTILOG with pre-processing and MULTILOG without pre-processing, are compared in terms of their performances on item parameter and student ability estimations with data from a popular physics concept test.

For parameter estimation, the three methods were shown to produce varying results. All methods produce very similar estimations of item difficulty. For item discrimination, R consistently produces higher estimates for all test items. The situation for the guessing parameter seems to be significantly affected by the pre-processing procedure, which imposes a Gaussian prior distribution for the guessing parameters, centered at 0.2 because of the five-choice multiple choice test. Without using pre-processing, near-zero guessing parameters are produced for 8 of the 30 items by both R and MULTILOG. With pre-processing, the guessing parameters on those 8 items increase. As shown by research in physics education, students coming into the course often have strong, naïve preconceptions, which can cause them choose responses based on incorrect conceptual thinking, resulting in scores close to zero, much lower than the theoretical chance level. Therefore, when implementing IRT analysis, we need to carefully inspect the assessment model against the cognitive models underlying the measurement instrument. This study shows an example suggesting that pre-processing might not be appropriate for use with certain conceptual test instruments, such as the FCI.

On the estimation of student ability, the differences among the three methods are on the order of 0.1, which is about 2 to 3% of the raw score difference. For a 30-question test, such differences are equivalent to the uncertainty of missing one test question, which is tolerable in education assessment, where we typically see standard deviations of 20% and effect sizes of about 0.5 (which is approximately equivalent to 50% of the standard deviation).

In terms of the goodness of fit, all methods seem to fit well, with slight variations. R matches the center average better, but has a larger range of variation, as it also stretches to cover the extremes of the data. The MULTILOG methods produce a more stable fit for different items with more weight on the center part of the data. However, MULTILOG consistently produces slightly higher estimates of student abilities than R does.

In summary, although there are variations, the compared IRT tools produce satisfying outcomes when analyzing a physics concept test. This study reveals interesting differences among the different methods, and these differences are important for researchers and teachers to consider when applying IRT methods in education assessment and interpreting the results of their analysis. In addition, examples in this study provide insight into the need for inspecting assessment models and adapting them to the cognitive models of the measurement instruments.

ACKNOWLEDGMENTS

The authors would like to thank all members of the Physics Education Research Group at The Ohio State University for their long-term support and valuable suggestions on this project. We would also like to acknowledge the help of the anonymous reviewers. The research is supported in part by NIH Award RC1RR028402 and NSF Awards DUE-0633473 and DUE-1044724. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of the National Institute of Health and the National Science Foundation.

ENDNOTES AND REFERENCES:

Corresponding Authors:

[1]Lei Bao, Physics department, The Ohio State University, 191 W.Woodruff Ave, PRB1016, Columbus, OH, 43210, USA

Phone: (614)292-2450 Fax: (614)292-7557

Email: bao.15@osu.edu

[2]Yan Tu, Dongfei Display R&D Center, Electronic Science and Engineering College, Southeast University, Nanjing, P.R. China 210096

Email: tuyan@seu.edu.cn

Bao, L., & Redish, E. F. (2006). Model Analysis: Assessing the Dynamics of Student Learning, Phys. Rev. ST Phys. Educ. Res. 2, 010103.

Bao, L., Hogg, K., & Zollman, D. (2002). Model Analysis of Fine Structures of Student Models: An Example with Newton's Third Law," Am. J. Phys. 70 (7), 766-778.

Demars, C. E. (2001). Group differences based on IRT scores: does the model matter?, Educational and Psychological Measurement, 60, 60-70.

Demars, C. E. (2002). Recovery of Graded Response and Partial Credit Parameters in MULTILOG and PARSCALE, Paper presented at the Annual Meeting of the American Educational Research Association.

Du Toit, M. (2003). IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT.

Hake, R. R., (1998). Interactive-engagement v.s. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. 66 (1): 64-74.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Norwell, MA: Kluwer Academic Publishers.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. The Physics Teacher, 30, 141-158. The version used in our analysis is the 1995 revised version.

Johnson, M. S. (2007). Marginal Maximum Likelihood Estimation of Item Response Models in R, Journal of Statistical Software, 20, 1-24.

Junker, B. W. (1999). Some statistical models and computational methods that may be useful for cognitively-relevant assessment, Paper prepared for the committee on the Foundations of Assessment, National Research Council.

Jurich, D., & Goodman, J. T. (2009). A comparison of IRT Parameter recovery in mixed format examinations using PARSCALE and ICL, Poster presented at the Annual meeting of Northeastern Educational Research Association.

Pellegrino, J. W. et al. (2001). Knowing What Students Know: The Science and Design of Educational Assessment.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications source. Journal of Educational Statistics, 4, 207-230.

Rizopoulos, D. (2006). ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses, Journal of Statistical Software, 17, 1-25.

Rizopoulos, D. (2009). Latent Trait Models under IRT,

http://wiki.r-project.org/rwiki/doku.php?id=packages:cran:ltm

Skaggs, G., & Stevenson, J. (1989). A comparison of pseudo-Bayesian and joint maximum likelihood procedures for estimating item parameters in the three-parameter IRT model, Applied psychological measurement, 13, 391-402.

Wang, J., & Bao, L. (2010). Analyzing force concept inventory with item response theory, Am.J.Phys, 78, 1064-1070.