Item response theory analysis of the mechanics baseline test

Citation: Cardamone, Caroline N. et al. “Item Response Theory Analysis of the Mechanics Baseline Test.” 2011 Physics Education Research Conference : Omaha, Nebraska, USA : 3-4 August 2011" editors, N. Sanjay Rebello, Paula V. Engelhardt, Chandralekha Singh., American Institute of Physics, Melville, N.Y.2012. (AIP Conf. Proc. p. 135–138). CrossRef. Web.

As Published: http://dx.doi.org/10.1063/1.3680012

Publisher: American Institute of Physics

Persistent URL: http://hdl.handle.net/1721.1/78319

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Item Response Theory Analysis of the Mechanics Baseline Test

Caroline N. Cardamone, Jonathan E. Abbott, Saif Rayyan, Daniel T. Seaton, Andrew Pawl[*], and David E. Pritchard

Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139.
[*]*Department of Chemistry and Engineering Physics, University of Wisconsin-Platteville, Platteville, WI 53818.*

Abstract. Item response theory is useful in both the development and evaluation of assessments and in computing standardized measures of student performance. In item response theory, individual parameters (difficulty, discrimination) for each item or question are fit by item response models. These parameters provide a means for evaluating a test and offer a better measure of student skill than a raw test score, because each skill calculation considers not only the number of questions answered correctly, but the individual properties of all questions answered. Here, we present the results from an analysis of the Mechanics Baseline Test given at MIT during 2005-2010. Using the item parameters, we identify questions on the Mechanics Baseline Test that are not effective in discriminating between MIT students of different abilities. We show that a limited subset of the highest quality questions on the Mechanics Baseline Test returns accurate measures of student skill. We compare student skills as determined by item response theory to the more traditional measurement of the raw score and show that a comparable measure of learning gain can be computed.

Keywords: physics education research, item response theory, testing
PACS: 01.40.Fk, 01.10.gf, 01.50.Kw

INTRODUCTION

Measurement of student learning and skill are fundamental components of physics education research. Traditionally, standardized pre and post-tests such as the Force Concept Inventory [1] and the Mechanics Baseline Test [MBT; 2] are used to evaluate and compare the effectiveness of instruction at many levels over diverse student populations [e.g., 3]. For these (or any) instruments, Item Response Theory (IRT) measures student skills better than total score, and provides insight into individual questions as we demonstrate by analyzing data from the MBT.

IRT provides a measure of the effectiveness and quality of each individual problem (or item) by identifying:

> **Item parameters**: *difficulty* identifies the absolute difficulty of an item, and *discrimination* determines how effective a given item is at distinguishing high and low skilled students.

Using the IRT fits we have identified two pathological items on the MBT that are more likely to be answered correctly by very unskilled students than by more skilled students. Clearly the presence of such items weakens the correlation of MBT score with student ability.

IRT determines standardized abilities from student responses to individual problems, rather than from a single total score on a standardized instrument or test [see e.g., 4, 5] by defining:

> **Student skill**: IRT optimally determines the student's skill from their responses to a set of problems with known item parameters. While in classical test scoring each correct response counts equally towards the total score, IRT skill is able to weigh different items differently, and the same item differently for different students.

For example, items with low discrimination are weighted lightly for everyone and a correct response to a item whose difficulty is comparable to a student's skill is weighted more heavily for that student because it provides more information about their skill.

A huge advantage of IRT is that student skill can be determined from any subset of calibrated items. Hence class skills can be determined at several times by administering only a subset of the MBT. Or total test time can be reduced with little sacrifice in accuracy by selecting items whose difficulty matches that of the class we wish to assess.

We present the IRT parameters for the MBT using a large set of data taken from pre and post tests given at MIT during 2005, 2007, 2008, 2009, and 2010 giving a total sample of 4754 tests. We first use these item parameters to evaluate the individual items in the MBT. We then show how a subset of the MBT items provides a measure of student skill comparable to the entire test. Finally, we compare IRT skills to classical test scores to

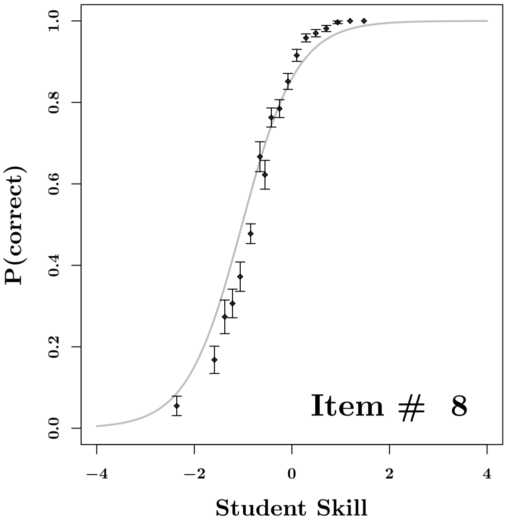

FIGURE 1. Example logistic item response function (thin dashed line, $\alpha = 1.8$, $\delta = -1.0$) expresses the probability that a student with a given skill level will answer the item correctly. Additional points and error-bars reflect the fraction of students (in bins of 50-300 students by IRT-determined skill level) that correctly answered this question.

TABLE 1. IRT parameters for the MBT.

Question	Difficulty		Discrimination	
1[1]	-4.80	± 0.76	**0.26**	± 0.04
2[1]	-4.90	± 0.80	**0.25**	± 0.04
3	-1.74	± 0.06	1.56	± 0.09
4[1]	-1.21	± 0.15	**0.35**	± 0.04
5	0.97	± 0.05	0.98	± 0.05
6[2]	-1.79	± 0.08	1.17	± 0.06
7[2]	-0.46	± 0.05	0.77	± 0.04
8[2]	-1.02	± 0.04	1.77	± 0.08
9	-0.22	± 0.04	1.11	± 0.05
10	-1.21	± 0.04	1.52	± 0.07
11	-0.45	± 0.03	1.46	± 0.06
12	1.07	± 0.05	1.18	± 0.06
13	-1.05	± 0.05	1.11	± 0.05
14[2]	-2.46	± 0.16	0.76	± 0.05
15	-1.92	± 0.08	1.23	± 0.07
16[2]	-0.81	± 0.03	1.52	± 0.07
17[1]	-2.55	± 0.28	**0.35**	± 0.04
18	1.27	± 0.08	0.76	± 0.04
19	-0.63	± 0.05	0.77	± 0.04
20	0.12	± 0.04	0.79	± 0.04
21	-2.53	± 0.15	0.94	± 0.07
22[1]	-0.62	± 0.11	**0.33**	± 0.04
23[2]	-2.83	± 0.23	0.54	± 0.05
24	-3.16	± 0.24	0.68	± 0.06
25	-1.61	± 0.11	0.60	± 0.04
26	-0.63	± 0.03	1.44	± 0.06

[1] Items with low discrimination.
[2] Items with a similar difficulty to a more discriminating item.

demonstrate that skills can be used to evaluate student learning in a manner comparable to pre and post testing.

ITEM RESPONSE THEORY

The key to IRT analysis is the item response function, which expresses the probability that a student of a given skill level (θ) will answer an item of difficulty (δ) and discrimination (α) correctly. It assumes a single continuous latent skill variable is being assessed by questions that are locally independent. Additionally, the item response function for an item i is assumed to be well represented by a logistic function:

$$P_i(\theta) = \frac{e^{\alpha_i(\theta-\delta_i)}}{1+e^{\alpha_i(\theta-\delta_i)}}. \qquad (1)$$

For example, Figure 1 plots the item response function for item #8 on the MBT. The points show the fraction of students correctly answering the question binned by skill level. The latent variable (shown on the x-axis), characterizes both the skill of the student and the difficulty of the question. The scale is set such that an average performing student has skill of 0, and the standard deviation of the student population is 1.

IRT PARAMETERS

After experimenting with several model fitting codes and finding consistent results, we use the open source package in R known as `ltm` for all results presented here [7]. Because of the large set of data, we are able to fit a two parameter logistic model, which fits individual discriminations α_i for each item to dichotomous data. The model considers each students answer to the multiple choice MBT exam as either correct, incorrect or blank if no response is given. The model parameters (item difficulty, item discrimination, and student skill) are estimated by marginalizing the maximum log-likelihoods of the observed data. MBT item parameters (and 1σ errors) are given in Table 1: note that questions with larger values of discrimination are better at distinguishing high and low skill students.

USING IRT TO ANALYZE THE MBT

The discrimination (α) determined by item response theory provides a useful measure by which items can be evaluated and better tests can be constructed [6, 8, 4]. In Table 1, there are 5 MBT items whose discrimination is significantly lower than average (items 1, 2, 4, 17, & 22). They are summarized in Table 2 and in this section, we take a closer look at each of these items.

Two of these items (# 1 and # 2) are not well matched to the MIT population's skill level. Over 70% of the test population answers these correctly; therefore, they do not provide a high level of score differential for students of different skills.

TABLE 2. MBT Items of Poor Discrimination

Items	Notes
1 & 2	Items too easy for student population
17	Students of all skill levels misread
4 & 22	Low skill students perform better than average

Question 17 refers to the acceleration of a car towing an object of twice its mass. Although 60% of the students answer this question correctly, nearly 30% select an answer indicating that they solved the problem using a total mass of twice the car's mass, forgetting to include the mass of the car in their computation. When we look at the distributions of the student skills, we find the average skill of a student selecting the correct answer ($\theta \sim 0.07$) is similar to the average skill of a student selecting the answer indicating they solved the problem with a total mass of twice the car ($\theta \sim -0.19$). Moreover, both answers were selected by a number of the students at the highest skill levels. Further study, including interviews, may suggest alternative wordings to this item that could improve its discrimination value.

There are two items (4 & 22) whose performance is not well represented by an item response function because low-skill students are more likely to answer the question correctly than students with average skill. Such behavior invalidates the assumption of classical testing theory that more skill results in higher overall scores. This is illustrated graphically by the "U-shape" in Figure 2, where it is apparent that the item response function is not a good fit to the data. Question 4 refers to the acceleration of a block sliding down a ramp with first a straight and then a curved section, just at the instant before the block enters the curved portion. A skilled student may misread the diagram, perceiving the portion of the track where the block is located to be curved, and hence getting this question incorrect. In contrast, students of low skill often confuse the concepts of acceleration and velocity and hence answer the question correctly. Question 22 refers to a diagram showing two pucks of different mass being pushed by equal forces. Students misinterpreting whether the force is impulsive or applied over a portion of time will answer this question incorrectly. Looking at their response patterns on the other two questions for this diagram, most students who misinterpret the force to be impulsive in question 22, also answer question 20 assuming that an impulsive force was applied.

In summary, the discrimination parameter identifies items that are ineffective in determining student skill. Additional physics education research can suggest ways to improve the performance of these items.

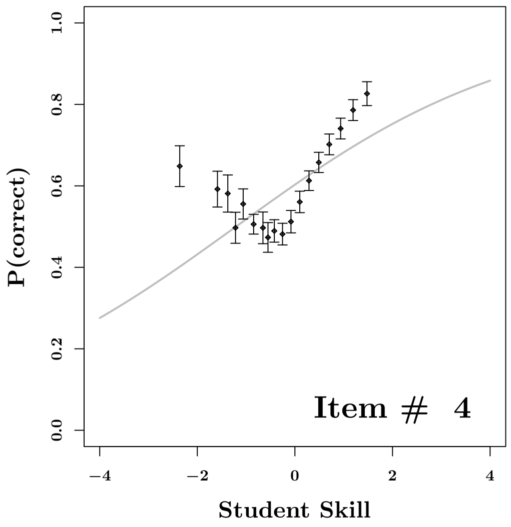

FIGURE 2. Logistic Item Response model for Question 4 (lines and symbols are as in Figure 1). The fraction of students answering each question correctly forms a strange "U-shape," representing the fact that students of the lowest skill levels were more likely to answer this question correctly than students whose skill level was average for the MBT.

IMPROVING THE TEST

Because an IRT skill is determined by the individual items, it is a more efficient measure of student performance. We show that IRT enables student skill to be computed equally well using a smaller set of items.

First, we remove the 5 questions with very low discrimination, and recompute the student skills using the remaining 21 items. Eliminating these questions changes the overall raw test score such that lower skill students appropriately get even lower relative scores on the shortened exam. Since IRT discounts the eliminated questions (due to their low discrimination) , there is a strong correlation between the skills determined using the full 26 question and the subset of 21 questions ($R = 0.996$). Therefore, we have improved the exam's ability to identify low skill students, making the resulting test score a better representative of the intrinsic student skill.

We next shorten the exam by eliminating items that are redundant. There are 6 pairs of items on the MBT exam that have similar difficulties (differences of less than 0.1). We remove the 6 paired items with lower discriminations (items 6, 7, 8, 14, 16, 23) and the 5 questions with low discrimination and recompute the students skills using the remaining 15 items. Again there is a strong correlation between the skills determined using the full 26 questions and the smaller exam with 15 items ($R = 0.97$). Therefore, we find that with IRT analysis we can create a more efficient assessment (15 questions instead of 26), and measure student skills nearly identical to those measured using the full MBT exam.

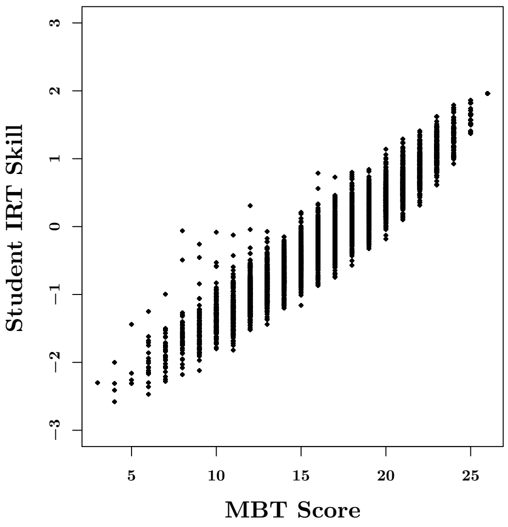

FIGURE 3. MBT Score (out of 26 questions) vs. Student IRT Skill as measured by the 2PL model. A single score value can be interpreted over a range of skills depending on the difficulties of the questions answered correctly.

IRT VS. CLASSICAL TEST THEORY

The skills determined through IRT can provide comparable measurements of student performance as the more commonly used raw test score. Figure 3 shows the comparison between the skills determined by IRT and the total MBT scores. The IRT skill is highly correlated with the classical test score ($R = 0.96$). However, the skills determined by IRT depend on the individual item parameters of the questions answered correctly and incorrectly. A blank response in IRT skill is not counted as wrong and IRT counts wrong answers as indicative of low skill. In contrast the total test score depends only on the number of items answered correctly. For example, a student with 15 correct responses on the MBT exam can have a skill from near -1 to 0 depending on which questions they answered correctly.

The average gain ($\% < \mathit{Post} > -\% < \mathit{Pre} >$) is a common measure of student learning between a pre and post test [3]. In Table 3, we compare the gain in IRT skill (the difference in the skill determined on the post and pre test on a scale where 1 is the standard deviation of the student population) to a more traditional measure of gain (the difference in the percent score on the post and pre test). The gain in IRT skill reflects the same gains seen in the percent correct on the pre and post test.

TABLE 3. Gain in IRT Skill and % Correct

Year	Skill Post - Skill Pre	$< \%\mathit{Post} > - < \%\mathit{Pre} >$
2005	0.38	9%
2007	0.90	20%
2008	0.98	21%
2009	0.86	17%
2010	0.82	18%

SUMMARY & FUTURE DIRECTIONS

By applying IRT analysis to the MBT exam we have shown that we can use the item parameters to identify items that do not effectively distinguish more from less skillful students. Furthermore, we have shown that we can use this item analysis to select a smaller subset of items to administer to a given student population while retaining or improving the accuracy of the measurement of the students skill. Finally, we have demonstrated that IRTs measurement of student skill can be used in the same way as classical tests scores to evaluate gains in learning.

We are now applying IRT to a preliminary analysis of online homework assignments presented as part of our Integrated Learning Environment for Mechanics [9]. Online systems contain the data necessary for the implementation of IRT analysis, providing an ideal environment in which a student's skill can be determined throughout the semester [10]. However, while IRT is well calibrated and tested in the domain of tests with only one possible response, extending this technique to assess student learning in an environment with multiple attempts is not as straightforward [10]. Studies are currently underway about how best to apply this technique in a domain where multiple responses are allowed to a variety of question types.

This work was supported by grants PHY-0757931 and DUE-1044294 from the NSF, which does not endorse this work.

REFERENCES

1. D. Hestenes, M. Wells, and G. Swackhamer, *Phys. Teach.* **30**, 141–158 (1992).
2. D. Hestenes, and M. Wells, *Phys. Teach.* **30**, 159–166 (1992).
3. R. R. Hake, *Am. J. Phys.* **66**, 64–74 (1998).
4. L. Ding, and R. Beichner, *Phys. Rev. ST Phys. Educ. Res.* **5**, 020103 (2009).
5. R. de Ayala, *The Theory and Practice of Item Response Theory*, The Guilford Press, New York, NY, 2009.
6. R. Hambleton, and R. Jones, *Educational Measurement: Issues and Practice* **12**, 38–47 (1993), ISSN 1745-3992, `http://dx.doi.org/10.1111/j.1745-3992.1993.tb00543.x`.
7. D. Rizopoulos, *Journal of Statistical Software* **17**, 1–25 (2006), ISSN 1548-7660, `http://www.jstatsoft.org/v17/i05`.
8. J. Buck, K. Wage, and M. Hjalmarson, *Item Response Analysis of the Continuous-Time Signals and Systems Concept Inventory*, in *Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, 2009. DSP/SPE 2009. IEEE 13th*, 2009, pp. 726 –730.
9. R. Teodorescu, A. Pawl, S. Rayyan, A. Barrantes, and D. Pritchard, *2010 Physics Education Research Conference Proceedings* (2010).
10. Y. Lee, D. Palazzo, R. Warnakulasooriya, and D. Pritchard, *Phys. Rev. ST Phys. Educ. Res.* **4**, 010102 (2008).