Item Response Theory Analysis
of the
Mechanics Baseline Test

C. Cardamone*, A. Barrantes, S. Rayyan, D. Seaton, R. E. Teodorescu, A. Wu & D. E. Pritchard

*cnc@mit.edu

Physics Department, MIT

http://relate.mit.edu

Abstract

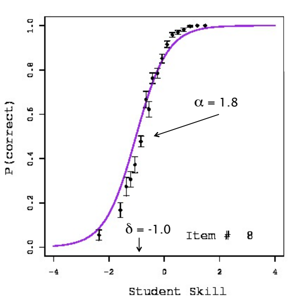
Item response theory (**IRT**) provides a more accurate measure of a student's skill than overall test score by employing a Bayesian calculation that considers the individual characteristics of each item (question) the student responds to, right or wrong. It also determines individual item parameters (difficulty, discrimination). In turn, these measure the effectiveness of the item in determining student skill, and identify items with pathological behavior (e.g. less skillful students outperform more skillful ones). These data allow evaluation and improvement of a test.

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 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.

Item Response Theory

The item response function expresses the probability that a student of a given skill level (θ) will answer an item of difficulty (δ) and discrimination (α) correctly.

The Item Response Function



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

δ: shifts an item left/right

α: slope of curve

θ: latent variable (student skill)

★ **Item Parameters:** accurate and efficient assessments can be designed with the most effective items.

- ★ *Difficulty* identifies items most useful in evaluating a given population of students.
- ★ *Discrimination* determines how effective a given item is at distinguishing high and low skilled students.

★ **Student skill:** depends on the difficulty and discrimination of each item and are more accurate than a classical test score, which depends only on the number of items correct.

Acknowledgements

This work was supported by grants PHY-0757931, DUE-1044294, NSF # 0757931 and NIH # 1RC1RR028302 - 01

The LearningOnline Network with CAPA

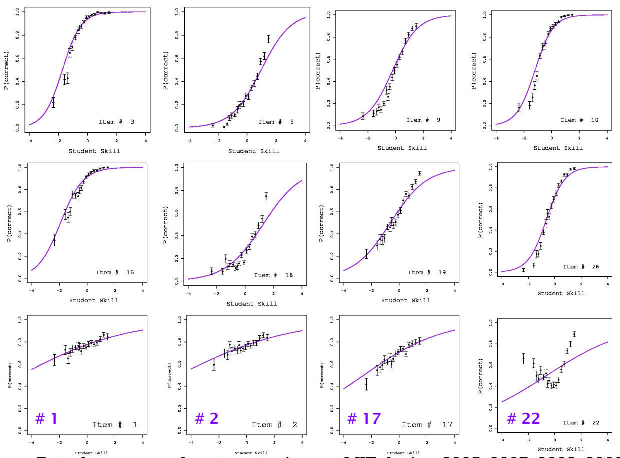


TABLE I. IRT parameters for the Mechanics Baseline Test.

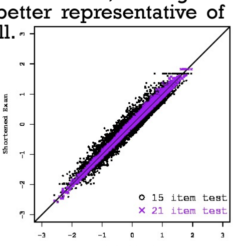
Question	Difficulty		Discrimination	
1[a]	-4.80	± 0.76	**0.26**	± 0.04
2[a]	-4.90	± 0.80	**0.25**	± 0.04
3	-1.74	± 0.06	1.56	± 0.09
4[a]	-1.21	± 0.15	**0.35**	± 0.04
5	0.97	± 0.05	0.98	± 0.05
6[b]	-1.79	± 0.08	1.17	± 0.06
7[b]	-0.46	± 0.05	0.77	± 0.04
8[b]	-1.02	± 0.04	1.77	± 0.08
9	-0.22	± 0.04	1.11	± 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[b]	-2.46	± 0.16	0.76	± 0.05
15	-1.92	± 0.08	1.23	± 0.07
16[b]	-0.81	± 0.03	1.52	± 0.07
17[a]	-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[a]	-0.62	± 0.11	**0.33**	± 0.04
23[b]	-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

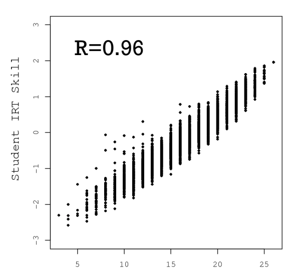
Data from pre and post tests given at MIT during 2005, 2007, 2008, 2009, & 2010 (a total sample of 4754 tests).

Improving the Test

IRT skill can be computed using a smaller set of items.

- 21 Qs: Remove 5 questions of low α (Table I "a"):
  - Raw test score: lower skill students get even lower relative scores on the shortened exam.
  - IRT Skills are unchanged (R=0.996), because it discounts the eliminated questions due to their low discrimination.
  - 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.
- 15 Qs: Also remove 6 redundant items (Table 1 "b"):
  - IRT skills are still the same as those determined by the full exam (R=0.97).

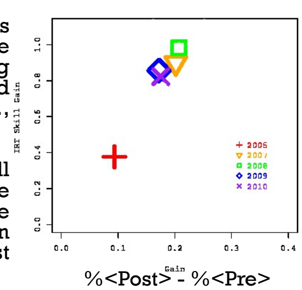




The skills determined by IRT depend on the individual item parameters of the questions answered correctly and incorrectly. In contrast the total test score depends only on the number of items answered correctly.

The average gain is a common measure of student learning between a pre and post test (e.g., Hake 1998).

The gain in IRT skill reflects the same gains seen in the percent correct on the pre and post test.



Items of Poor δ

★ Item 1 & 2: too easy for the student population

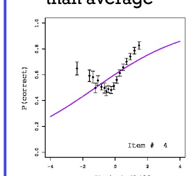
★ Item 17: students of all skill levels misread

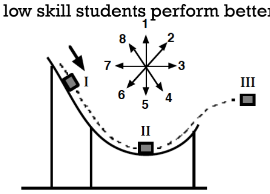
17. A car has a maximum acceleration of 3.0 m/s^2. What would its maximum acceleration be while towing a second car twice its mass?
(A) 2.5 m/s^2 (B) 2.0 m/s^2 (C) 1.5 m/s^2 (D) 1.0 m/s^2 (E) 0.5 m/s^2

- 60 % of the students answer correctly (average skill θ = 0.07)
- 30 % select answer c, forgetting to include the mass of the first car in their computation (average skill θ = -0.19)
- Students at the highest skill levels selected both answers

Further study, including interviews, may suggest alternative wordings to this item that could improve its discrimination value.

★ Item 4 & 22: low skill students perform better than average



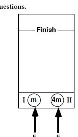


4) Indicate the direction of acceleration at I

Skilled students may perceive the track where the block is located to be curved and select "2". In contrast, students of low skill often confuse the concepts of acceleration and velocity and hence select "4".

Refer to the diagram below when answering the next three questions.

The diagram depicts two pucks on a frictionless table. Puck II is four times as massive as puck I. Starting from rest, the pucks are pushed across the table by two **equal** forces.



20. Which puck will have the greater kinetic energy upon reaching the finish line?
(A) I (B) II
(C) They both have the same amount.
(D) Too little information to answer.

21. Which puck will reach the finish line first?
(A) I (B) II
(C) They will both reach the finish line at the same time.
(D) Too little information to answer.

22. Which puck will have the greater momentum upon reaching the finish line?
(A) I (B) II
(C) They will both have the same momentum.
(D) Too little information to answer.

Students misinterpreting the force as impulsive rather than constant in 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.

References

- R. R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," American Journal of Physics, 66, 64–74 (1998).