All submissions to PERS should be sent (preferably electronically) to the Editorial Office of AJP, and then they will be forwarded to the PERS editor for consideration.

Testing the test: Item response curves and test quality

Gary A. Morris
Department of Physics and Astronomy, Valparaiso University, Valparaiso, Indiana 46383

Lee Branum-Martin
Texas Institute for Measurement, Evaluation, and Statistics, University of Houston, Houston, Texas 77204

Nathan Harshman
Department of Computer Science, Audio Technology, and Physics, American University, Washington, DC 20016

Stephen D. Baker
Department of Physics and Astronomy, Rice University, Houston, Texas 77251

Eric Mazur and Suvendra Dutta
Division of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138

Taha Mzoughi
Department of Physics and Astronomy, Mississippi State University, Mississippi 39762 and College of Science and Mathematics, Kennesaw State University, Kennesaw, Georgia 30144

Veronica McCauley
National University of Ireland, Galway, Ireland

(Received 19 September 2005; accepted 20 January 2006)

We present a simple technique for evaluating multiple-choice questions and their answers beyond the usual measures of difficulty and the effectiveness of distractors. The technique involves the construction and qualitative consideration of item response curves and is based on item response theory from the field of education measurement. To demonstrate the technique, we apply item response curve analysis to three questions from the Force Concept Inventory. Item response curve analysis allows us to characterize qualitatively whether these questions are efficient, where efficient is defined in terms of the construction, performance, and discrimination of a question and its answer choices. This technique can be used to develop future multiple-choice examination questions and a better understanding of results from existing diagnostic instruments. © *2006 American Association of Physics Teachers.*

[DOI: 10.1119/1.2174053]

I. INTRODUCTION

Physics education research has contributed significantly to the processes of evaluation, criticism, and innovation in physics teaching.[1] Assessment is the engine of this change; assessment tools inform researchers as to whether new teaching methods improve the knowledge, skills, and attitudes of the students. Physics education research requires a solid understanding of the performance and effectiveness of the testing assessment instruments. For multiple-choice instruments we would like to develop better questions and better combinations of questions.

In this paper we use item response curve (IRC) analysis to evaluate the quality of individual multiple-choice questions and answer choices that are on the Force Concept Inventory (FCI),[2] a diagnostic assessment instrument for kinematics and Newtonian mechanics. Item response curves relate the percentage of students at each ability level to each possible answer choice. Figure 1 shows a set of item response curves for the five answer choices from a question on the FCI. Both the question and the construction of these curves are discussed in Secs. II and III. IRC analysis can be applied to any multiple-choice assessment.

Significant research into student misconceptions of physics concepts has been conducted.[3–9] As a result, the questions on standard conceptual diagnostic examinations such as the FCI typically contain distractors that are based on common misconceptions. Therefore, we may expect that even incorrect answers to these questions can tell us about the level of student knowledge and understanding.

The quality of a question commonly has been evaluated using simple statistical measures described in education measurement theory.[10,11] Two of these measures are the question difficulty and the effectiveness of the distractors. The question difficulty is related to the percentage of students who do not select the correct answer choice. The effectiveness of a distractor is measured by the percentage of students selecting a particular incorrect answer choice.

Although measuring the difficulty of a question and the effectiveness of its distractors provides important insights,

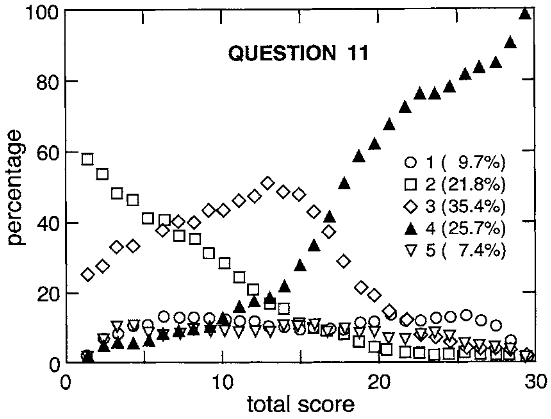


Fig. 1. Item response curves (IRCs) for Question 11 from the Force Concept Inventory. The percentage of students who select each particular answer choice is plotted versus total score. In this case, the correct answer choice is 4.

our IRC analysis, which is a simplified version of item response theory,[12] leads to further understanding. For instance, the IRC analysis provides insight into the discrimination of an answer choice: the sharper the change in the percentage of students who select a given answer choice as a function of student ability, the more discriminating the answer choice is. To convey this point, imagine an extreme (and unlikely) case in which a question yields a step-function-like IRC for the correct answer. Such an IRC would be ideally discriminating: all students below a given ability level fail to select the correct answer choice, while all students above that ability level do select the correct answer choice. Although quantitative measures of discrimination can be generated using item response theory, we limit ourselves here to a qualitative interpretation of the discrimination of answer choices as indicated by the IRCs. By examining the discrimination of the answer choices, instructors might gain insight into student knowledge and understanding through an analysis of incorrect responses to a difficult question (one that very few students get right).

Figure 1 shows an example using actual data. The IRCs for several answers have distinct increases or declines with ability level (for example, answer choice 4), while others have nearly constant appeal over the entire ability range (answer choice 1). (In this study we used the student's total score on the FCI as a proxy for ability level.) We would say that answer choice 4 is more discriminating than answer choice 1. Unlike the standard dichotomous scoring, which only discriminates on the basis of selecting or not selecting the correct answer, the IRC analysis permits a richer evaluation of student understanding based on all of his/her answer choices.

Once the performance of the questions and specific answer choices has been evaluated, better test questions can be developed. Answer choices that are ignored by students or are nondiscriminating can be either eliminated or replaced with revised answer choices. Further, an IRC analysis can indicate that in some cases, entire questions require significant revision or elimination.

Carrying out these methods of question analysis and evaluation allows us not only to make better use of existing tests to assess student understanding, but also to construct better tests that are more sensitive to different student ability levels. As Beichner[4] has pointed out, constructing a diagnostic test is an iterative procedure. We believe that the IRC analysis can become a powerful tool as part of this process and improve our ability to develop and analyze questions and answer choices designed to test specific concepts. As with student interviews, IRC analysis can lead to the development of questions that contain only proven, functional distractors. As a result, IRC analysis may allow us to develop a test bank from which to construct conceptual examinations that are uncluttered by poor questions or answer choices, thereby providing a mechanism for a more thorough and fair evaluation of our students.

An approach built on item response theory (from which IRC analysis is derived) might also lead to new ways of scoring, with different incorrect answer choices given different weights. We also hope that IRC analysis will facilitate the identification and the investigation of multiple-choice test questions as research topics for future physics education research projects.

Below we briefly describe the IRC analysis approach and then discuss our results for three illustrative questions on the FCI that are, respectively, "efficient," "moderately efficient," and "inefficient." By efficient, we mean that the question has the following characteristics: (1) the answer choices are discriminating (i.e., they distinguish among student ability levels); (2) different answer choices reflect different conceptual or physical models; and (3) each answer choice is selected by some significant number of students (or, if not, at least serves a logical purpose). In the cases we discuss, the IRC analysis provides insight beyond the standard dichotomous-scoring approach.

We note that there are formally defined statistical methods for evaluating the difficulty and discrimination of multiple-choice test items.[10,11] However, many of these statistical approaches require specialized training and software. Examples of such approaches are found in Ref. 13 and in the nominal response model,[14] the latter of which is implemented in MULTILOG (a program available through Scientific Software International). Our intention is to demonstrate how the IRC analysis provides a simple and qualitative approach for evaluating multiple-choice test responses. Formal treatments of statistical models for multiple-choice tests can be found in Refs. 14–17. General introductions to item response theory can be found in Refs. 10–12 and 18.

II. ITEM RESPONSE CURVE ANALYSIS

To perform an IRC analysis we need a database containing item-level responses for each student. It is not sufficient to know simply whether the student answered a particular question correctly or incorrectly. IRC analysis requires a knowledge of which specific answer choice a student selected for each item on the test. IRC analysis also requires a measure of the ability level of the student. Frequently, the total score on the examination can serve as a proxy for this variable. Student scores on the Scholastic Aptitude Test or the American College Testing exam also could serve in this regard. Additional data such as gender, age, year-of-study, and prior coursework could be included in the analysis. We will not include such additional variables here, although an extension of the IRC analysis to these variables is straightforward. For this study we gathered FCI data from students at Harvard University, Mississippi State University, and Rice University,

Table I. Percentage of students responding to each of the five possible answer choices on three questions from the Force Concept Inventory. The percentages for the correct answer choices are in bold type.

Answer choice	Question No.		
	Q4	Q9	Q11
1	59.6	4.5	9.7
2	1.1	21.9	21.8
3	0.9	20.0	35.4
4	1.3	7.8	**25.7**
5	**37.1**	**45.9**	7.4

and we used the total score on the FCI as a proxy for the student ability level. The combined data set includes more than 4500 individual examinations.

For each question on the FCI we determined the percentage of students at each ability level who selected each of the five answer choices. Scores on the FCI can range from 0 to 30 so there are 31 different possible outcomes or ability levels. Because only one student in our data set scored a 0 on the test, we will use the 30 levels ranging from 1 to 30. For smaller data sets we might choose to bin the scores.

After performing this analysis item-by-item, question-by-question, and ability-level-by-ability-level, we produced a graph for each question with the percentage response as the dependent variable and total score as the independent variable. Each graph contains five curves, one for each of the answer choices.

III. EXAMPLES USING FCI DATA

A. An efficient question

FCI Question 11 probes student understanding of the nature of impulse and Newton's First Law (see the Appendix for the question statement). The question is difficult as only 25.7% of the students answered correctly. Table I lists the percentage of students who selected each of the five answer choices. Such a percentage analysis provides little information beyond that the distractors are selected by significant numbers of students.

Figure 1 shows the IRCs for Question 11, one for each answer choice, and provides more revealing information than the percentage analysis summarized in Table I. The curve for answer choice 4, the correct answer, shows that the probability of selecting this answer choice is correlated with ability level. Furthermore, the shape of the curve shows that this answer choice discriminates strongly between the lower ability and higher ability students, with a rapid rise in the IRC around an ability level of 17. Such sharp discrimination is one feature of an efficient answer choice.

Our IRC analysis indicates that answer choices 2 and 3 are discriminating. Answer choice 2 is the most popular choice of students of very low ability level. The misconception represented by this answer appears to be present in students of lowest ability. In fact, few students with a total score of $\geqslant 22$ selected this answer choice. Answer choice 3 is the most popular choice of students in the middle to low range of ability with total scores between 4 and 17. Again, this answer choice discriminates strongly between students of high ability, who avoid it, and those of middle to low ability, who prefer it. The other two answer choices, 1 and 5, do not discriminate and seem to be equally attractive to students of nearly all ability levels. One possible explanation for these IRCs is that students selected these answer choices randomly (through guessing).

Let us examine the content of answer choices 2 and 3 further to understand why the selection of answer choice 3 indicates a deeper understanding by the student than the selection of answer choice 2. Both answer choices 2 and 3 contain the erroneous suggestion that a force is required in the direction of motion. They also both include the force of gravity, which is part of the correct answer. What differentiates the two answer choices is the inclusion in answer choice 3 of the contact force of the table, a force that is also part of the correct answer choice. From an analysis of the physics content of these two answers, we would conclude that answer choice 3 is preferable to answer choice 2; that is, its content is closer to the correct answer. The IRC analysis concurs with our expectations based on the physics content: students who selected answer choice 3 probably have mastered the material somewhat better than those who selected answer choice 2 because they recognized the need for the contact force directed opposite to the force of gravity to prevent vertical acceleration of the puck. Although we could have reached such a conclusion without the IRC analysis, a statistical basis for our judgments about the differential quality of multiple-choice answers promotes a more nuanced and defensible evaluation of our students.

Students often lament that multiple-choice questions offer no partial credit. The IRC analysis permits us to evaluate our students not only on the basis of their correct responses, but also on the basis of their incorrect answers. Some incorrect answer choices really do reflect higher student ability levels than others, as demonstrated by the IRC analysis of Question 11 shown in Fig. 1. This analysis raises the possibility of assigning partial credit for some of the incorrect answers, a main goal of other, more detailed multiple-choice item-response models (for example, Ref. 14).

Although we classify this question as efficient, judgments on the appropriateness of the question and its physics content need to be made separately by content experts. For example, we note that Question 11 asks the students to base their answers on their answers to Question 8. It is logically possible for a student who answers Question 8 incorrectly to select a response to Question 11 that, although consistent with his/her answer to Question 8, would be considered an incorrect response to Question 11. The IRC and other statistical analyses would probably not have identified this issue, so expert interpretation and analysis remain essential.

B. Moderately efficient question

Our IRC analysis of the responses to Question 9 from the FCI shows it to be moderately efficient. Although it describes the same physical situation as Question 11, Question 9 probes the student's understanding of the effect of an impulse delivered to a moving object on its velocity vector (see the Appendix). This question is moderately difficult, with 45.9% selecting the correct answer choice 5. The data in Table I also suggest that perhaps better distractors could be found to replace answer choices 1 and 4, which attract significantly smaller percentages of students than the other two incorrect answers.

Figure 2 shows the IRC analysis graph for this question. Two of the distractors (answer choices 1 and 4) not only

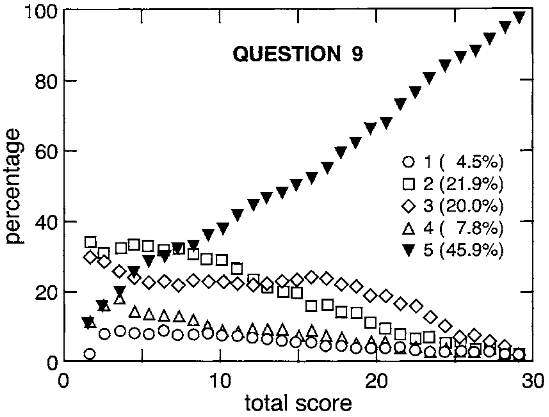


Fig. 2. As in Fig. 1, but for Question 9 from the FCI. The correct answer choice is 5.

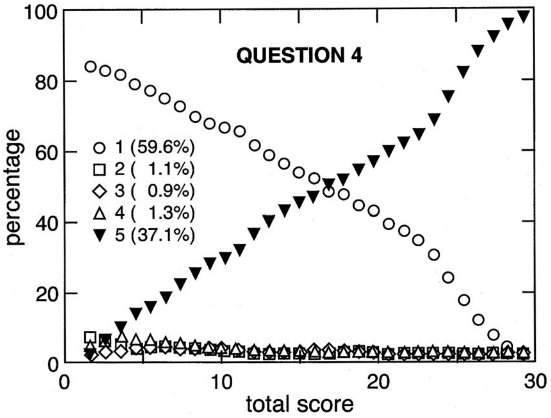


Fig. 3. As in Fig. 1, but for Question 4 from the FCI. The correct answer choice is 5.

attract few students but also are not discriminating. The correct answer choice 5 is also not discriminating. Although the IRC for answer choice 5 indicates that the higher the ability of the student, the greater the likelihood of the student's selecting the correct answer, its linear nature indicates that it discriminates no better than the overall score on the test. Thus answer choice 5 does not yield an IRC that would be helpful in identifying students of different ability levels. Answer choices 2 and 3 both attract a reasonable percentage of the students and are somewhat more discriminating that the correct answer choice 5. Answer choice 2 is somewhat more attractive to the lowest ability students, but the slope of the curve for students with total scores of 5–25 is almost constant. Answer choice 3, the most discriminating of the choices with a change in the slope of the IRC occurring in the total score range 20–25, appears to appeal to a slightly more capable cohort of students than answer choice 2. The IRCs for the incorrect answer choices 2 and 3 may thus be more informative regarding student ability than the IRC for the correct answer choice 5.

Let us examine the content of answer choices 2 and 3 further. Answer choice 2 ignores the initial momentum of the puck, asserting that only the speed that would result from the kick matters in the final velocity of the puck. Answer choice 3 recognizes that the initial momentum of the puck must be combined with the change in momentum due to the kick, but adds the two velocities together arithmetically rather than vectorially. Which of these errors is the less forgivable is difficult to say. It is not surprising, therefore, that the IRC curves are similar. In fact, students who selected answer choice 2 may have been confused about the statement of the problem, especially if they failed to carefully read the paragraph describing the physical situation that appears prior to the statement of Questions 8–11: in particular, what is meant by "the speed v_k resulting from the kick?" Does v_k include the contribution due to the initial momentum of the puck?

The student responses to Question 9 are more difficult to evaluate than Question 11. Whether or not Question 9 could be modified so as to become more efficient requires investigation through student interviews and the testing of new distractors, an investigation well worth pursuing but beyond the scope of this paper.

C. An inefficient question

Question 4 from the FCI tests student understanding of forces in collisions (see the Appendix). As seen from Table I, only 37.1% of the students choose the correct answer 5, indicating that this question is not easy. Answers choice 2 (car exerts a bigger force), 3 (neither exerts a force), and 4 (truck exerts a force but car does not exert a force) are each selected by approximately 1% of the students.

Figure 3 shows the IRC analysis graph for Question 4. We see that all of the IRCs, including those for the two most popular answer choices, 1 (truck exerts a greater force) and 5 (forces are equal, the correct answer) are not discriminating: the IRCs for these choices are nearly linear over the entire range of total scores. Figure 3 suggests that this question could be reworked into a true/false question because students overwhelmingly selected answer choices 1 and 5 and the IRCs for answer choices 2, 3, and 4 are flat. The high correlation of the correct answer with student ability level indicates that the presence of this question on this multiple-choice test is not informative.

IV. SUMMARY

We have introduced IRC analysis and demonstrated the technique through an analysis of student responses to questions from the FCI. IRC analysis provides a powerful new tool for developing and analyzing multiple-choice questions such as those found on diagnostic exams. In the future an even richer analysis of multiple-choice questions can be achieved using a fuller application of item response theory than the IRC analysis presented here. The IRC approach can provide instructors with a method for scoring multiple-choice questions beyond the standard dichotomous scheme of correct/incorrect—some wrong answers really are better than others.

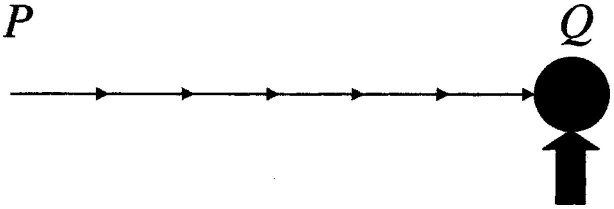
APPENDIX: FORCE CONCEPT INVENTORY QUESTIONS

We show the selected questions from the Force Concept Inventory[2] discussed in this paper, reproduced here with permission. The student responses and our analysis are included for questions 4, 9, and 11. Question 8 is included for completeness because Question 11 refers to the content of Question 8.

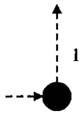
4. A large truck collides head-on with a small compact car. During the collision

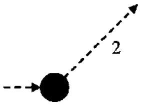
(1) the truck exerts a greater amount of force on the car than the car exerts on the truck.
(2) the car exerts a greater amount of force on the truck than the truck exerts on the car.
(3) neither exerts a force on the other, the car gets smashed simply because it gets in the way of the truck.
(4) the truck exerts a force on the car but the car does not exert a force on the truck.
(5) the truck exerts the same amount of force on the car as the car exerts on the truck.

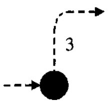
Use the statement and diagram below to answer the next four questions (8–11). The diagram depicts a hockey puck sliding with constant speed v_0 in a straight line from point P to point Q on a frictionless horizontal surface. Forces exerted by the air are negligible. You are looking down on the puck. When the puck reaches point Q, it receives a swift horizontal kick in the direction of the heavy print arrow. Had the puck been at rest at point P, then the kick would have set the puck in horizontal motion with a speed v_k in the direction of the kick.

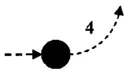


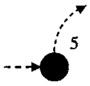
8. Which of the paths 1–5 below would the puck most closely follow after receiving the kick?











9. The speed of the puck just after it receives the kick is

(1) equal to the speed v_0 it had before it received the kick.
(2) equal to the speed v_k resulting from the kick and independent of the speed v_0.
(3) equal to the arithmetic sum of the speeds v_0 and v_k.
(4) smaller than either of the speeds v_0 or v_k.
(5) greater than either of the speeds v_0 or v_k, but less than the arithmetic sum of these two speeds.

11. Along the frictionless path you have chosen in Question 8, the main force(s) acting on the puck after receiving the kick is (are)

(1) a downward force of gravity.
(2) a downward force of gravity, and a horizontal force in the direction of motion.
(3) a downward force of gravity, an upward force exerted by the surface, and a horizontal force in the direction of motion.
(4) a downward force of gravity and an upward force exerted by the surface.
(5) none. (No forces act on the puck.)

[1] L. C. McDermott and E. F. Redish, "Resource Letter: PER-1: Physics education research," Am. J. Phys. **67**, 64–74 (1999).

[2] D. Hestenes, M. Wells, and G. Swackhamer, "Force concept inventory," Phys. Teach. **30**, 141–158 (1992).

[3] D. Hestenes and M. Wells, "A mechanics baseline test," Phys. Teach. **30**, 159–166 (1992).

[4] R. J. Beichner, "Testing student interpretation of kinematic graphs," Am. J. Phys. **62**, 750–762 (1994).

[5] R. K. Thornton and D. R. Sokoloff, "Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula," Am. J. Phys. **66**, 338–352 (1998).

[6] S. Yao and M. Zadnak, "Introductory thermal concept evaluation: Assessing students″ understanding," Phys. Teach. **39**, 496–504 (2001).

[7] D. P. Maloney, T. L. O'Kuma, C. Hieggelke, and A. Van Heuvelen, "Surveying students' conceptual knowledge of electricity and magnetism," Am. J. Phys. **69**, S12–S23 (2001).

[8] C. Singh and D. Rosengrant, "Multiple-choice test of energy and momentum concepts," Am. J. Phys. **71**, 607–617 (2003).

[9] P. Vetter Englehart and R. J. Beichner, "Students' understanding of direct current resistive electrical circuits," Am. J. Phys. **72**, 98–115 (2004).

[10] H. K. Suen, *Principles of Test Theories* (Lawrence Erlbaum Associates, Hillsdale, NJ, 1990).

[11] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Holt Reinhard and Winston, New York, 1986).

[12] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of Item Response Theory* (Sage Publications, Thousand Oaks, CA, 1991).

[13] F. Drasgow, M. V. Levine, S. Tsien, B. Williams, and A. D. Mead, "Fitting polytomous item response theory models to multiple-choice tests," Appl. Psychol. Meas. **19**, 143–165 (1995).

[14] R. D. Bock, "Estimating item parameters and latent ability when responses are scored in two or more nominal categories," Psychometrika **37**, 29–51 (1972).

[15] F. Samejima, A new family of models for the multiple-choice item, No. 79-4, University of Tennessee, Knoxville, TN, 1979.

[16] D. Thissen and L. Steinberg, "A response model for multiple choice items," Psychometrika **49**, 501–519 (1984).

[17] D. Thissen and L. Steinberg, "A taxonomy of item response models," Psychometrika **51**, 567–577 (1986).

[18] R. K. Hambleton, "Principles and selected applications of item response theory," in *Educational Measurement*, 3rd ed., edited by R. L. Linn (Macmillan, New York, 1989), pp. 147–200.