Making Sense of Students' Answers to Multiple-Choice Questions

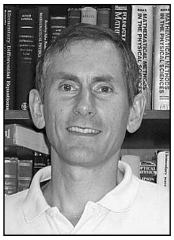
Robert J. Dufresne, William J. Leonard, and William J. Gerace

Multiple-choice questions (MCQs) are widely used in physics instruction. They are employed routinely to assess student understanding and achievement,[1] and increasingly, they are used in educational research to evaluate curriculum materials and teaching practices.[2] MCQs are popular because the multiple-choice format is easy to implement, and the grading and scoring of student answers can be automated, saving time and reducing costs. As a research mechanism the multiple-choice format provides an efficient way to collect and analyze data from large numbers of students.

Despite the merits and the widespread use of MCQs, many instructional applications in physics are not supported by independent research. Further, current models of student knowledge and problem solving in the domain of physics are often inadequate for making sense of students' answers to MCQs. In this article, a detailed example is used to illustrate the difficulties in making sense of students' answers to MCQs. We explore how correct answers can be false indicators of student knowledge and understanding. We suggest that care is needed when interpreting students' responses to MCQs, even if the MCQs have undergone a formal process of construction and validation.

What Is at Issue?

In order to base instructional decisions on students' responses to a set of MCQs, one needs to know how students' answers are related to what they actually know and understand. Students' answers are, at best, only suggestive of their underlying knowledge. Determining the precise relationship between students' answers to particular MCQs and what they actually comprehend requires correlating their answers with other measures of the same or related knowledge and understanding. For example, a question that is intended to



Robert J. Dufresne

Robert J. Dufresne is a Research Assistant Professor in the Department of Physics at the University of Massachusetts. He received his Ph.D. in Nuclear Theory from UMass in 1987. His research interests include assessment, instructional technology, and models of cognition. He is the coordinator of the Assessing-to-Learn materials development and teacher enhancement project.

Department of Physics, Box 34525
University of Massachusetts
Amherst, MA 01003–4525;
dufresne@physics.umass.edu

measure "student understanding of Newton's third law" must be investigated to determine what other types of understanding account for actual performance on the question.

We will use the phrase *response model* to describe the relationships between students' answers on a particular MCQ (or set of MCQs) to other indicators of their knowledge and understanding, and to contextual factors (such as experience, confidence, and mathematical ability) that could determine how they answer. (For illustration, a portion of a hypothetical response model is shown in Fig. 1.) A response model is critical to making sense of students' answers to MCQs and therefore an important resource for drawing conclusions about what students know — and by inference, about the value of a given instructional approach or set of curriculum materials. Unfortunately, forming a response model is difficult and time consuming, and without a theoretical basis to guide its construction, it is usually too unwieldy to be of practical use.

- → 80% of high school seniors choose the correct answer.
- → In interviews, 60% of students who have no high school physics get the correct answer by analogy with typical situations covered in ninth-grade physical science without showing any awareness or understanding of Newton's third law in other contexts;
- → 10% of students who have finished a year of physics guess the correct answer;
- → 40% of students who have finished a year of physics and answer the original MCQ correctly cannot answer correctly if the context is altered so that one of the objects has a very large mass or is moving very fast;
- → Only 15% of students who answer correctly are confident in their answer.

→
→
→

Fig. 1. Hypothetical response model for fictitious Newton's third law multiple-choice question.

What Can Go Wrong: An Example

The following example illustrates the need for a response model when making sense of students' answers to MCQs. In the example, we compare a sample of students' answers on a chosen MCQ with their answers on two complementary MCQs, all of which target related conceptual knowledge. The pattern of performance across the three questions suggests that a correct answer on the chosen MCQ is, *more often than not*, a false indicator of deep conceptual understanding (or, more precisely, a false indicator of deep conceptual understanding in the context under study).

For the chosen MCQ, we selected question 1 from the original Force Concept Inventory (FCI), a multiple-choice test that was designed to measure students' conceptual understanding of the Newtonian force concept.[3] Our decision to focus on an MCQ from the FCI was due to the perceived status of FCI items and their widespread recognition. They are based on educational research and target physics concepts that are difficult to learn. They are the products of an extensive development process for which the reliability and criterion-related validity of items has been checked. Finally, because they have been widely used, there exists a significant quantity of published data.[2,3]

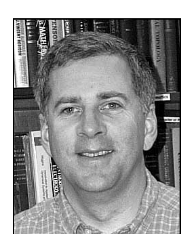
Survey

To investigate students' answers to question 1 on the FCI (FCI#1), we created a survey containing FCI#1 and two complementary multiple-choice questions (CMCQ#1 and CMCQ#2). All three MCQs were constructed around the same problem situation and are presented in Fig. 2.

In FCI#1 two balls of different weight are dropped

William J. Leonard is a Research Assistant Professor of Physics at the University of Massachusetts. He received his Ph.D. in Nuclear Theory from UMass in 1988. His research interests include teacher preparation, communication, and applying educational research to develop classroom materials. He is the coordinator of the Minds-On Physics materials development project.

Department of Physics, Box 34525
University of Massachusetts
Amherst, MA 01003–4525;
wjleonard@physics.umass.edu



William J. Leonard

from the same height, and students are asked to compare the times it takes for the two balls to reach the ground. FCI#1 is intended to probe students' conceptual understanding of the (local) gravitational force. Specifically, it probes the relationship between an object's weight and its acceleration — i.e., acceleration is independent of the object's weight.[3]

The two CMCQs were also intended to probe students' conceptual understanding of the gravitational force. In CMCQ#1 two balls having different weights are dropped from the same height, and students are asked to compare the forces exerted on the two balls. In CMCQ#2 students are asked to compare both the forces exerted on two objects and the times it takes for them to reach the ground.

The three questions are part of a larger survey, the details of which are described elsewhere.[4] Four different versions of the survey were created (see Table I), though for the present context there are only two subpopulations of students. Approximately half of the students received FCI#1 and CMCQ#2; the other half received

The following questions have been developed to determine your perceptions of real-life situations. Do not be concerned with getting the "correct" answer. We really want to know what you think is happening in these situations. The results of this questionnaire will **not** affect your semester grade in any way.

For each question, indicate the *Confidence Level* of your response:
1 = not at all confident; I guessed
2 = not very confident
3 = reasonably confident
4 = very, very confident

FCI#1. Two metal balls are the same size, but one weighs twice as much as the other. The balls are dropped from the top of a two-story building at the same instant of time. The time it takes for the balls to reach the ground below will be:
(A) about half as long for the heavier ball.
(B) about half as long for the lighter ball.
(C) about the same time for both balls.
(D) considerably less for the heavier ball, but not necessarily half as long.
(E) considerably less for the lighter ball, but not necessarily half as long.

CONFIDENCE LEVEL: (I guessed) 1 2 3 4 (very, very confident)

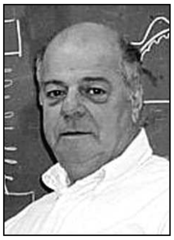
CMCQ#1. Two metal balls are the same size, but one weighs twice as much as the other. The balls are dropped from the top of a two-story building at the same instant of time. As the balls are falling, the force on the two balls is:
(A) about twice as large for the heavier ball.
(B) about twice as large for the lighter ball.
(C) about equal for both balls.
(D) considerably more for the heavier ball, but not necessarily twice as large.
(E) considerably more for the lighter ball, but not necessarily twice as large.

CONFIDENCE LEVEL: (I guessed) 1 2 3 4 (very, very confident)

CMCQ#2. Two equal-sized objects, one weighing 2 lbs and the other weighing 4 lbs, are released from rest from the roof of a two-story building. Which of the following statements is true?
(A) The force on the 4-lb object is about twice as large as the force on the 2-lb object, therefore, the 4-lb object reaches the ground in about half the time.
(B) The forces on the two objects are about equal, therefore, they both reach the ground at about the same time.
(C) The force on the 4-lb object is about twice as large as the force on the 2-lb object, but they both reach the ground at about the same time.
(D) The forces on the two objects are about equal, but the 4-lb object reaches the ground in about half the time.
(E) None of the above.

CONFIDENCE LEVEL: (I guessed) 1 2 3 4 (very, very confident)

Fig. 2. Survey instructions and three multiple-choice questions used to probe student understanding of the (local) gravitational force.


William J. Gerace

William J. Gerace is Professor of Physics and the Director of the Scientific Reasoning Research Institute at the University of Massachusetts. He received his B.S. from M.I.T. and his Ph.D. from Princeton University. His research interests include developing a cognitive model for the acquisition, storage, and use of knowledge, and teacher preparation.

Department of Physics, Box 34525
University of Massachusetts
Amherst, MA 01003–4525;
gerace@physics.umass.edu

Table I. Structure of the four versions of the survey.

Survey Question	Versions 1&2	Versions 3&4
1	FCI#1 (time)	CMCQ#1 (force)
4	CMCQ#2 (both force & time)	

CMCQ#1 and CMCQ#2. Either FCI#1 or CMCQ#1 is the first question on each student's survey, and CMCQ#2 is always the fourth and last question. (The other two questions are not relevant to this discussion.) In addition to answering each question, students were asked to indicate their level of confidence in each answer (see instructions in Fig. 2).

Sample and Procedure

The survey was administered at the University of Massachusetts and was given to two groups of students: (1) those enrolled in the introductory calculus-based mechanics course for engineering majors (Physics 151), and (2) those enrolled in the algebra-based introductory physics course for life-science majors (Physics 131). We administered the survey to five classes in all (three classes of Physics 151 and two classes of Physics 131) in three separate semesters over a period from spring 1995 to spring 1996. The survey was administered on the first or second day of class. Each student received one of the versions, selected randomly. Students worked individually, and they spent between 10 and 15 minutes completing the survey. The total number of students who participated is 1046.

Results and Discussion

The response patterns for the three questions, FCI#1, CMCQ#1, and CMCQ#2, are shown in Table II. The upper half of the table shows the breakdown of responses on the first question (FCI#1) by question 4 (CMCQ#2). The lower half of the table shows the breakdown of responses on the first question (CMCQ#1) by question 4 (CMCQ#2).

Answers to FCI#1. The upper half of the column on the extreme right shows the results for FCI#1. Approximately 69% of the students (362 out of 527) chose the correct answer [(C) the time is about the same for both balls]. The most frequently chosen incorrect answers [(A) and (D)] were those for which the heavier object takes less time to reach the ground. This result is consistent with research on students' common-sense beliefs, which suggests the existence of a common-sense concept: *heavier objects fall faster*. This response pattern for FCI#1 is in good agreement with previously published results.[3]

Multiple interpretations. There are many possible circumstances that could cause students to answer FCI#1 correctly without an appropriate understanding of Newtonian concepts. One possibility is that students guess the correct answer. Another possibility is that students know as a factual matter that *all objects fall at the same rate*; these students need not have acquired any understanding of Newtonian concepts such as force, gravitational force, weight, or acceleration. Still another possibility is that students have acquired some Newtonian force concepts, but have not made links between the different concepts. These students could know as a fact that all bodies fall at the same rate. They could also know that the two objects have the same acceleration and that the acceleration is 9.8 m/s^2. They could have a firm grasp of the concept of acceleration. Nevertheless, it is quite possible for students to know all of this (and much more) without knowing why, from a Newtonian viewpoint, two objects of different weight have the same acceleration in a local gravitational field.

Answers to CMCQ#1. Students' answers to CMCQ#1 (see the lower half of the last column of Table II) suggest that many of the correct responses to FCI#1 are not based on a deep understanding of Newtonian force concepts. For CMCQ#1 only about 35% (180 out of 519) of students select the correct answer that *the force is about twice as large for the heavier ball*. This is only about half the rate of correct responses observed for FCI#1. Further, for CMCQ#1 nearly half of the students answering (256 out of 519) chose the incorrect answer (C): *the force on the two balls is about equal for both balls*.

The answer pattern for CMCQ#1 suggests that the rate of false positive answers is nearly 50% for FCI#1 (specifically, a false positive rate of nearly 50% in the current context). When the answer patterns for FCI#1 and CMCQ#1 are compared, the results are somewhat surprising in light of the research literature.

Table II. Results of first question (either time or force) vs fourth question (both time and force). A yellow background indicates a correct answer on only one of the questions, and a dark red background indicates a correct answer on both.

		Answer on Question 4, CMCQ#2 (force and time)						
Variation of Question 1	Answer on Question 1	(A) $F_4 \approx 2F_2$ $t_4 \approx ½t_2$	(B) $F_4 \approx F_2$ $t_4 \approx t_2$	(C) $F_4 \approx 2F_2$ $t_4 \approx t_2$	(D) $F_4 \approx F_2$ $t_4 \approx ½t_2$	(E) None of the above	(left blank)	TOTAL
FC#1 (time)	(A) $t_H \approx ½t_L$	20	1*	5*	18	0	2	46
(N = 527)	(B) $t_L \approx ½t_H$	9*	0*	2*	7*	0	0	18
	(C) $t_H \approx t_L$	5*	160	169	16*	5	7	362
	(D) $t_H << t_L$	26	2*	10*	32	5	3	78
	(E) $t_L << t_H$	3*	3*	4*	7*	1	1	19
	(left blank)	0	0	0	2	0	2	4
	TOTAL	63	166	190	82	11	15	527
CMCQ#1 (force)	(A) $F_H \approx 2F_L$	39	16*	111	8*	3	3	180
(N = 519)	(B) $F_L \approx 2F_H$	1*	1*	7*	2*	0	0	11
	(C) $F_H \approx F_L$	4*	160	41*	44	3	4	256
	(D) $F_H >> F_L$	15	6*	29	8*	5	2	65
	(E) $F_L >> F_H$	1*	0*	2*	0*	0	0	3
	(left blank)	0	0	1	0	0	3	4
	TOTAL	60	183	191	62	11	12	519

*These combinations of answers are *incompatible* with each other

For FCI#1, the percentage of correct answers is fairly high, even though one of the incorrect answers corresponds to a well-documented common-sense misconception: *heavier objects fall faster.*[3] For CMCQ#1, the percentage of correct answers is comparatively low, even though the correct answer corresponds to a possible common-sense conception: *larger weight means larger force*. Conversely, the percentage of students selecting the incorrect choice (C) for CMCQ#1 is relatively high, even though there is no corresponding common-sense misconception reported in the literature that *force is independent of weight.*

The described pattern of answers suggests the possibility that students are applying an over generalized rule: *for dropped objects, "things" are the same.* The "things" are the time in FCI#1 and the force in CMCQ#1. Since in CMCQ#1 no common-sense misconception appears to be associated with the popular incorrect response (C), and the correct answer (A) is associated with a common-sense concept, we conjecture that the over generalized rule results from students' rote learning in the science classroom. In other words, students have learned through drill-and-practice that the times are the same without learning the underlying conceptual basis, and therefore, students conclude that other features of the situation are the same as well.

Another possibility is that students still believe that the time is proportional to the force (i.e., *heavier objects fall faster*). Perhaps when they are told or witness that objects reach the ground at the same time, one way to accommodate this new information is to conclude that because the times are the same, the forces must be the same as well.

Answers to CMCQ#2. FCI#1 and CMCQ#1 were given to different, but we assume equivalent, subpopulations of students. CMCQ#2 was given to all students and asks them to compare both the forces on the balls and the times for the balls to reach the ground. The percentage of students selecting the correct answer to CMCQ#2 is about 36%. This is about half the rate at which students answer FCI#1 correctly (69%), but comparable to the rate at which students answer CMCQ#1 correctly (35%).

In aggregate, students' answers to CMCQ#2 appear consistent with their answers to FCI#1 and CMCQ#1. The rate at which students select an answer indicating that the balls reach the ground at the same time [CMCQ#2 answers (B) or (C)] is 70%, compared with 69% in FCI#1 [answer (C)]. The rate at which students select an answer indicating that the force on the heavier object is larger [CMCQ#2 answers (A) or (C)] is about 48%, compared with 47% on CMCQ#1 [answers (A) and (D)]. Finally, the rate at which students select an answer indicating that the force is the same on both objects [CMCQ#2 answers (B) or (D)] is about 47%, compared with 49% on CMCQ#1 [answer (C)].

In summary, for FCI#1 more than two-thirds of the students indicated that the times for the two balls to reach the ground are the same. Based on students' answers to CMCQ#1 and CMCQ#2, we learn that only a little more than a quarter of the students consistently indicated that the times are the same for both balls *and* that the force is about twice as large on the heavier ball. These results suggest that as few as a quarter of the students may actually understand the situation presented in FCI#1 in terms of Newtonian force concepts. If so, the use of FCI#1 as an indicator of the Newtonian force concept would lead to a false positive result (in the context under study) approximately 60% of the time.

Compatibility. Most students' combinations of answers are *compatible* with each other. For instance, of the 362 students who answered in FCI#1 that the times are equal [answer (C)], 91% also chose that the times are the same in CMCQ#2 [answers (B) or (C)]. Overall, about 84% of students chose compatible pairs of answers.

The other 16% of students (171 of 1046), however, are a bit of a mystery. On a four-question survey, how can someone, for instance, answer that the forces are about equal on question 1, then answer that the forces are different on question 4? Some of these incompatibilities were forced upon students by the structure of an early version of the survey (22 of 171), because they were not given the option of (E) *None of the above* on CMCQ#2, but the rest are not forced.

The largest group of inconsistent students (45 of them) answered that the forces on the two balls are about equal for CMCQ#1, but that the force on the heavier ball is larger for CMCQ#2. The next largest group (38) did the reverse, answering that the force on the heavier ball is larger for CMCQ#1 and that the forces are equal for CMCQ#2. One possible interpretation of these results is that students were not taking the survey seriously, which means that their answers on different questions would not be related to each other. This possibility cannot be easily ruled out, complicating efforts to make sense of students' answers. Another possibility is that students answer question 1 without any interference from the complementary idea. Then, upon encountering the two ideas together in question 4, they changed their views without going back and changing their answers to question 1.

The incompatibility rates for the two subpopulations are not the same: 11% of students in the FCI#1 subgroup chose an incompatible answer on CMCQ#2 (60 out of 527), while 17% of the CMCQ#1 subgroup did (89 out of 519), which means that it is about 50% more likely that a student who answered the force question first will choose an incompatible combination of answers. Our conjecture is that thinking about the times (in question 4) induces a change in their views about force more often than thinking about the force induces a change in their views about the time, perhaps because students are more certain in their views about time.

Confidence levels. Generally students were confident in their answers, presumably indicating that most students were not simply guessing. Overall, 83% of students were confident in their answers to FCI#1 (indicating a confidence level of 3 or 4). Only 74% of these confident students, however, answer the question correctly.

A slightly lower percentage of students (76%) were confident in their answer to CMCQ#1. A much smaller fraction of the confident students (36%) responded correctly (i.e., *the force on the heavier object is twice as large*). Dismayingly, 52% of confident students answered incorrectly that *the forces are about equal*, and among all students who answer that *the forces are about equal*, 80% were confident in their answers.

On CMCQ#2, which asks students to compare both the forces on the two balls and the times it takes the two balls to reach the ground, the percent of students who were confident in their answers dropped to 64%, and of these, only 41% answered correctly. Among the students who answered correctly, 74%

were confident. However, nearly the same fraction (72%) of students who answered incorrectly *that the forces are about equal, so the balls fall at the same rate* were also confident.

For those students applying a general rule in rote fashion, we expect approximately the same level of confidence on each of the questions — i.e., the same level of confidence independent of the quantity being compared. For students not using a general rule, we would expect the confidence level to vary greatly with the student's familiarity with, and understanding of, both the quantity and the problem situation.[5] Among students who selected an answer indicating that *the forces are about equal* on both questions, we find that the average confidence level is approximately the same (3.2 for CMCQ#1 and 3.1 for CMCQ#2). Interestingly, the number of confident students is roughly the same for both questions (132 out of 160 for CMCQ#1 and 120 out of 160 for CMCQ#2).

Some Final Comments

Through an example, we have attempted to show that making sense of students' answers to even well-designed MCQs that are used by a large number of educators is an inherently difficult task. We have tried to highlight how correct answers in some contexts can be false indicators of students' knowledge and understanding. A detailed response model is helpful to adequately interpret students' answers to MCQs and use these answers to make sound teaching decisions. More research, however, is needed to understand the relationships between students' answers to multiple-choice questions, students' knowledge and understanding, and the contextual factors that could influence the way students answer questions.

Although MCQs have merit in physics instruction, some caution is warranted in their use. It is unlikely that any fixed set of MCQs can adequately represent students' knowledge. Information about student knowledge and understanding — as well as students' reasoning abilities and problem-solving skills — should be sought from a variety of sources. These other sources of information might include student writings, oral discussions and presentations, problem-solving projects, and Socratic dialogue. It is only through multiple sources of information that a reliable assessment of a student's knowledge and understanding can be obtained.

References

1. MCQs are also used as in-class learning activities for students. See, for example, Robert Dufresne et al., "*Classtalk:* A classroom communication system for active learning," *J. Comp. High. Ed.* 7(2), 3 (Spring 1996); Laura Wenk et al., "Technology-assisted active learning in large lectures," in *Student-active Science: Models of Innovation in College Science Teaching*, edited by Ann P. McNeal and Charlene D'Avanzo (Saunders, Orlando, FL, 1997), pp. 431–451; and Eric Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Upper Saddle River, NJ, 1997).
2. Richard R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *Am. J. Phys.* **66**(1), 64 (1998); online at http://www.physics.indiana.edu/~sdi/.
3. For the original *TPT* article introducing the FCI, see David Hestenes, Malcolm Wells, and Gregg Swackhamer, "Force concept inventory," *Phys. Teach.* **30**(3), 159 (1992). For the latest version of the FCI, see I. Halloun et al., Force Concept Inventory (Revised, 1995) password protected at http://modeling.la.asu.edu/modeling.html. Note: The results reported here concern one MCQ from the FCI and are intended to illustrate through an example the nature of the difficulties interpreting students' answers to MCQs. The results cannot be generalized to the entire FCI exam, nor do they speak directly to the validity of various applications employing the FCI.
4. The complete survey can be found online at http://umperg.physics.umass.edu/writings/supplements/makingSense.
5. This relationship between confidence and performance would not hold if a widespread misconception were involved, since a misconception would lead to an incorrect answer with an artificially high confidence level. The results of this study suggest that there might be a widely held alternative conception that the forces are about equal for two falling objects even though they have different weights, with as many as one-third of students with the equivalent of two or more years of high school physics holding this view. This is consistent with results previously found for 13-year-olds [Varda Bar et al., "Children's concepts about weight and free fall," *Sci. Ed.* **78**(2), 149 (1994)] and high school physics students [James Minstrell, "Facets of students' thinking," on the web at http://depts.washington.edu/huntlab/diagnoser/facetcode.html#300; also David Brown and John Clement, "Classroom teaching experiments in mechanics," in *Research in Physics Learning: Theoretical Issues and Empirical Studies*, edited by Reinders Duit, Fred Goldberg, and Hans Neidderer (IPN, Kiel, Germany, 1992), pp. 380–397].