Evaluating multiple-choice exams in large introductory physics courses

Michael Scott, Tim Stelzer, and Gary Gladding

Department of Physics, University of Illinois at Urbana-Champaign, 1110 W. Green St., Urbana, Illinois 61801, USA

(Received 11 May 2005; published 28 July 2006)

The reliability and validity of professionally written multiple-choice exams have been extensively studied for exams such as the SAT, graduate record examination, and the force concept inventory. Much of the success of these multiple-choice exams is attributed to the careful construction of each question, as well as each response. In this study, the reliability and validity of scores from multiple-choice exams written for and administered in the large introductory physics courses at the University of Illinois, Urbana-Champaign were investigated. The reliability of exam scores over the course of a semester results in approximately a 3% uncertainty in students' total semester exam score. This semester test score uncertainty yields an uncertainty in the students' assigned letter grade that is less than $\frac{1}{3}$ of a letter grade. To study the validity of exam scores, a subset of students were ranked independently based on their multiple-choice score, graded explanations, and student interviews. The ranking of these students based on their multiple-choice score was found to be consistent with the ranking assigned by physics instructors based on the students' written explanations ($r > 0.94$ at the 95% confidence level) and oral interviews ($r = 0.94^{+0.06}_{-0.09}$).

DOI: 10.1103/PhysRevSTPER.2.020102

PACS number(s): 01.40.Fk

I. INTRODUCTION

The Department of Physics at the University of Illinois, Urbana-Champaign began reforming its introductory physics sequence in the fall of 1996.¹ As part of the reform, midterm and final exams were converted from constructed-response to multiple-choice format. Prior to this reform, the physics exams had been relatively traditional exams in which students were asked to solve problems and were given credit based on the correctness of their written work. With classes as large as 1000 students, grading the exams and assigning partial credit in a consistent manner was a major endeavor. Even with trained graders using rubrics, inconsistencies arise among different graders as well as for a given grader between different students. Students often felt the allocation of partial credit was unfair, and a significant amount of time was spent dealing with student appeals. This likely produced further systematic effects as outspoken students were more likely to succeed in getting their exams regraded. The net effect of this exam format was that both professors and students were frustrated by the experience.

The difficulty of reliably grading large numbers of exams is not unique to physics and has been extensively studied by professional testing agencies. Much of the research has focused on comparing the multiple-choice format with the constructed-response format. Lukhele *et al.* from the educational testing service found that, on a chemistry advanced placement (AP) examination, "a 75 min multiple-choice test is as reliable as a 185 min test built of constructed-response questions."² In the time to give a single-constructed response question, they could give many more multiple-choice questions and receive more information about the students. They also found that "to predict a particular student's score on a future test made up of constructed-response items," they "could do so more accurately from a multiple choice than from a constructed-response test that took the same amount of examinee time." Hence, many of the national exams such as AP exams and the graduate record examination (GRE) utilize the multiple-choice format.

Switching to the multiple-choice format solved the grading difficulties experienced with the constructed-response exams. Student complaints about grading essentially disappeared, with the occasional exception being exam questions that could legitimately be open to multiple interpretations. Still there remained considerable concern about the ability of multiple-choice exams to accurately assess students' understanding.^{3,4} Although significant research has been performed for professionally constructed exams, there is little or no research that exists on the validity or reliability of multiple-choice exams constructed by course instructors. Indeed, much of the success of the national exams is attributed to the careful construction and testing of each item to ensure its effectiveness. This procedure is unrealistic in physics departments where exams are generally created in a short period of time by one or more members of the faculty who have little or no formal training in exam construction. The goal of this study was to determine if multiple-choice exams created in the Department of Physics at the University of Illinois yield scores that are reliable and valid assessments of student understanding in introductory physics. A discussion of the construction and evaluation of the multiple-choice exams is given in Appendix A. To see all of the midterm, multiple-choice exams used in the introductory courses in recent years, visit the Illinois Physics Education Research Group's web site at <http://www.physics.uiuc.edu/Research/PER/> and click on the "Resources" link.

Exam construction experts measure the ability of an exam to assess student understanding based on the reliability and validity of the exam scores. Reliability refers to the reproducibility of students' scores, i.e., the extent to which one would expect a student's score to vary if the student was given another equivalent exam. Validity refers to the extent to which exam scores are representative of what the writer intends to measure. Reliability and validity are two dimensions that can be used to evaluate scores from an examination. Exam scores can be reliable, but not valid, if they are measured precisely and are repeatable but are not indicative of what one wants to measure, i.e., the scores are not accu-

FIG. 1. (Color) Percent difference in the students' even and odd test scores. Items were split based on item difficulty.

rate measurements. Exam scores can also be valid while not being reliable if they are measured accurately with what the instructor intends but there is a large amount of uncertainty in each measurement, i.e., the scores are not very precise.

Section II of this paper describes two methods for determining the reliability of exam scores and how this can be used to estimate the uncertainty in a student's exam score. Section III describes the study that was conducted to determine the validity of exams scores from one of the multiple-choice exams. Grading students' written work and how it can be implemented into a course apart from exams is discussed in Sec. IV. In Sec. V is a summary of the work and results presented.

II. RELIABILITY

Reliability is the extent to which a student's exam results are reproducible. To estimate the reliability of exam scores, students can be given two similar exams, both in content and in difficulty.⁵ The distribution of the differences between the two sets of test scores for each student provides one estimate for quantifying the reliability of exam scores. A narrow distribution in the set of test score differences would suggest that the exam results are reliable, whereas a broad distribution would suggest that student exam scores are not reproducible.

Ideally to perform this type of analysis, one would administer two separate but equivalent exams to each student throughout the semester. This, however, is not practical. Instead, one can take the complete set of exam items and split them into two equivalent sets, e.g., split by even and odd numbered questions or split by item difficulty.

A split-exam analysis was used to determine students' semester exam score uncertainties for all four introductory physics courses at the University of Illinois at Urbana-Champaign. Students in these courses take 4 multiple-choice exams each semester: 3 midterms and a final. Each midterm consists of 25 to 30 questions and the final exam is approximately 50 questions.⁶ To make two equivalent exams, calling them euphemistically "even" and "odd" exams, the semester set of exam items were split based on item difficulty.⁷ Figure 1 shows the cumulative results for both the algebra and

FIG. 2. (Color) Graph of students' average total semester test score binned by their grade point average. Data are taken from the calculus-based E&M course from the 2003 spring semester.

calculus-based courses between the years 1999 and 2003. This combines 32 different course semesters, 128 multiple-choice exams, 4250 questions, and 12 281 students (A+ to C- only).^{8,9} A Gaussian fit to the data reveals a 3.1% standard deviation based on this splitting method. Other splitting methods give similar results.^{10–12}

This result is consistent with an error estimate based on a binomial distribution:¹³

$$\text{percent error} = \sqrt{\frac{p(1-p)}{N}}, \quad (1)$$

where p is the average test score and N is the number of test questions. For our 32 courses, the average test score is 73% and the average number of questions in a semester is 133 giving an estimated percent uncertainty in a student's semester test score of 3.85%.

The uncertainty in student scores is one important measure for quantifying the reliability of exam scores. However, this uncertainty must be normalized with the standard deviation of the class scores to obtain an estimate of reliability. For example, a 3% uncertainty in a student's semester test score would not give much information about that student if there were only a 3% difference separating the "A" students from the "D" students.

When course grades are essentially dependent upon exam performance, a letter grade uncertainty can then be estimated from the student's exam score uncertainty. Figure 2 shows the correlation between exam score and course grade for students in an introductory physics course. Translating the uncertainty in exam score into a letter grade uncertainty (l.g.u.) is achieved by dividing the test score uncertainty by the slope of the best-fit line of the average total test score versus grade point average:

$$\text{l.g.u.} = \frac{\text{total test score uncertainty}}{\text{slope of avg. total test score vs gpa}}. \quad (2)$$

When this is done for the 32 different course semesters of our introductory courses, the letter grade uncertainty ranges from $\frac{1}{4}$ to $\frac{1}{3}$ of a letter grade, with the average uncertainty being 0.27.¹⁴ Here, the mapping of letter grades to grade point average is an A=4.0; an A-=3.7; a B+=3.3; a B=3.0; etc. Therefore, the letter grade uncertainties reported here are less than the difference between a letter grade of an

TABLE I. Estimated semester test reliability coefficients along with their corresponding test score uncertainties for the three methods of splitting the semester's set of exam questions. These values are for A+ to C- students only.

Splitting method	r_{exam}	Exam score uncertainty
No. 1	0.896 ± 0.003	$3.17\% \pm 0.04\%$
No. 2	0.896 ± 0.003	$3.18\% \pm 0.04\%$
No. 3	0.888 ± 0.003	$3.29\% \pm 0.05\%$

A and an A- or the difference between an A- and a B+.

Mathematically using true score theory,¹⁵ this test score uncertainty can be understood using the correlation coefficient between the even and odd split tests, $r_{e,o}$. In brief, a student's exam score uncertainty can be found using Eq. (3):¹⁶

$$\text{exam score uncertainty} = \sigma_{exam} \sqrt{1 - r_{exam}}. \quad (3)$$

Here σ_{exam} is the standard deviation in exam scores and r_{exam} is the reliability coefficient for the exam. An exam's reliability coefficient tells how correlated students' scores would be between that exam and a similar exam. The value of r_{exam} can be estimated by splitting the exam into two equivalent sets (e.g., even and odd questions) and using the correlation between these two split halves, $r_{e,o}$. The exam's reliability coefficient can then be found using the Spearman-Brown formula¹⁶

$$r_{exam} = \frac{2r_{e,o}}{1 + r_{e,o}}. \quad (4)$$

Table I lists semester test reliability coefficients along with their predicted exam score uncertainty for the three methods of splitting the exam questions (A+ to C--students only).¹⁷

Rather than providing justifications for specific splitting methods, one can take the conservative approach of estimating the reliability in terms of the Cronbach alpha, α , which is the average of all exam reliability coefficients that could be obtained from the different splittings of a test:¹⁸

$$\alpha = \frac{N}{N-1} \left(1 - \frac{\sum \sigma_i^2}{\sigma^2} \right). \quad (5)$$

Here N is the number of test questions, σ_i^2 is the variance in scores for the i th question, and σ^2 is the variance in the total test scores.¹⁹ The average exam reliability coefficient for the 32 course semesters using Eq. (5) was 0.87 ± 0.01 , which leads to an average semester test score uncertainty of $3.5\% \pm 0.1\%$. This uncertainty is consistent with the value obtained using the specific split-exam methods. Figure 3 shows a histogram of alpha coefficients in each of the 32 semester courses for all students (A to F).

III. VALIDITY

Determining the reliability of exam scores is relatively straightforward. Assessing the validity of exam scores is

FIG. 3. (Color) A histogram of alpha coefficients for the semester set of exams for each of the 32 courses for all students (A–F).

much more difficult as it requires comparing a student's exam results with an assessment of the student's physics knowledge. This comparison can be even more difficult for a multiple-choice format exam. It is well known that scores on multiple-choice format questions can depend strongly on the nature of the distractors, or even the position of the correct answer in the list.³ Exams used in our introductory courses are used to assess the relative level of mastery of the specified curriculum, with the scores ultimately being used to assign letter grades to students. The goal of this study is to compare the scores students receive from a multiple-choice exam to those they would receive from a constructed-response exam. It is, therefore, instructive to know whether the distribution of assigned letter grades to students is consistent across exam formats. We use the scores from constructed-response exams, where the student's work can be examined by physics instructors, as the assessment of the students's physics knowledge. This procedure for assessing the validity of multiple-choice tests was developed early in the 20th.²⁰ We attempt to improve this assessment of the student's physics knowledge by supplementing the written constructed-response questions with an interview with the student designed to help clarify any ambiguities that remained after reviewing the written work. The details of this study are presented below.

A. The study

In the spring of 1999, two similar, multiple-choice finals were given for the introductory electricity and magnetism (E&M) course for physics and engineering majors. Of the two populations of students who took the final exams, a select subset of students who scored consistently on their first three midterms were invited to participate in the validity study and received \$20 for their participation. Roughly equal numbers of students who received A's, B's, and C's on the first three midterm exams were accepted. This selection process was chosen to ensure a uniform distribution of student abilities for which the exam is designed to differentiate.²¹ The number of students in the subset of those who took the first final was $N_1=16$ and of those who took the second final was $N_2=17$. In total there were 33 students who participated in the study which was 9% of the total number of students enrolled in the course.

Immediately after completing the multiple-choice course final exam, the students were taken to another room. In this

room they were asked to work through 20 questions selected from their final exam, this time showing all of their work. These questions covered five of the major topics discussed in this E&M course: electric fields, electric potential, Gauss' law, Coulomb's law, and Faraday's law. The students were allowed to see their final exam and use any notes they had made during the actual exam in completing this section. They also had the liberty to mark different answers on this 20 question follow-up form than they had marked originally during their multiple-choice test. Ideally it may have been better for the students to first complete the constructed-response portion first and then complete the multiple-choice exam due to the possibility of students' written work being influenced by the item choices present in each problem. However, this study was intended not to interfere with the structure of the course. Thus, the constructed-response exam was completed second.

Once each student completed the follow-up form, the student was interviewed by one of four physics instructors participating in the study. The interviewer reviewed the student's work and asked questions to assess the student's understanding of the material. Each interview had a duration of 10–20 min and was recorded onto audiocassettes.

The student's written explanations for each question were then independently graded by the same four physics instructors. The assigned grade was made on an integer scale between zero and three, with zero representing little or no knowledge of the physics involved in the problem and three representing full knowledge of the physics knowledge involved. The partial knowledge decision between a grade of one and two was made by determining whether or not credit would have been given had the grading been only credit or noncredit.

Once the independent grading had been completed, the instructors met in a committee to assign a grade to each question for each student. The objective of the committee was to assign grades based only upon the level of the student's understanding of the relevant physics. The committee based its score upon the independently graded scores, the recorded interviews, and the observations made by the interviewer. The committee also gave an integer score from 0 to 3 to each question for each student as described in the previous paragraph.

The 20 items from each final exam that were used in the validity study are listed in Appendix C. Full credit for a two-choice, three-choice, or five-choice question was two points, three points, or six points, respectively.²² To account for this weighted system, the scores assigned to the students by each grader and the committee were also weighted to have a parallel structure with the weights assigned to the different types of multiple-choice questions. In total, then, there are three sets of scores for the students: (1) their multiple-choice score (MC) from their original final exam, (2) their average-grader score from the four instructors (AG), and (3) their committee score (CS). The results in Part B of this section will address correlations between the MC and the other two sets of scores. Large correlation coefficients imply that students' MC scores are consistent with their scores from their graded solutions and their interviews.

TABLE II. Correlations between students' MC score and their AG score. These results imply that the students' multiple-choice score are indeed consistent with their scores given by instructors grading their written solutions.

Correlations between MC and AG		$N_1=16$	$N_2=17$
Raw correlation	$r_{MC,AG_{raw}}$	0.88	0.92
Corrected for attenuation	$r_{MC,AG_{atten}}$	1.00	1.00

B. Validity results

The raw correlations between the MC scores and the AG scores for the two groups were 0.88 and 0.92.²³ The probability, or p value, of obtaining these correlation values randomly is $p < 10^{-11}$. Because the study only involved probing students with 20 questions, there is a statistical correction to the correlation coefficient that is made to predict what the correlation would be had they answered an infinite number of questions. To know this correlation between MC and AG, we must correct our raw correlations for attenuation:^{16,20,24}

$$r_{MC,AG_{atten}} = \frac{r_{MC,AG_{raw}}}{\sqrt{r_{MC}r_{AG}}}. \quad (6)$$

Here r_{MC} and r_{AG} are the reliabilities of the multiple-choice questions and the average-grader scores, respectively, as explained in the Reliability section of this paper. Table II lists the different correlation values between MC and AG obtained for the two samples of students.

Similar studies were performed to compare students' MC scores with their CS scores. Raw correlations between the two sets of scores were 0.78 and 0.83 for the first and second group,²⁵ respectively, with a probability of randomly occurring at $p < 10^{-7}$. Correcting for attenuation raises the correlation values to 0.91 and 1.00.

C. Validity sensitivity

To determine the sensitivity of these validity results, we performed a Monte Carlo analysis. In particular, simulations were run for different assumed values of the true correlation coefficients to determine the probabilities of the observed correlation coefficients to fluctuate as high as those found in our validity study. This analysis then can be used to set lower limits on the true correlation coefficients between the MC and the AG (or CS) scores.

In a given simulation, the MC scores are first generated according to the observed MC score distribution from our data. For each MC score an AG (or CS) score is then generated using the assumed true value of the correlation coefficient, r_{true} , and the observed AG (or CS) score distribution from our data. A single run generates a set of 33 pairs of observed MC and AG (or CS) scores. From this single run, an observed correlation coefficient, r_{obs} , can be calculated. We then repeat this process thousands of times to create a distribution of the observed correlation coefficient, r_{obs} , that are generated by a specific r_{true} . From this distribution, we can calculate the probabilities to use in a maximum likelihood analysis.²⁶

FIG. 4. (Color) These are maximum likelihood plots for the combined groups of students generated from the simulation data to determine the true correlation coefficient between students' MC and AG scores and the MC and CS scores.

Figure 4 shows the maximum-likelihood fit using input values from the AG and CS scores. At the 95% confidence level, the true correlation r_{true} between students' MC score and their AG score is found to be greater than 0.94. When using input data from the CS scores, the true correlation between MC and CS was found to be $0.94^{+0.06}_{-0.09}$.

These results are encouraging and suggest that for our population of students, their MC scores are as valid as scores from constructed-response exams and oral interviews. However, there remains the possibility that a subpopulation that was not involved in the study may exist for whom MC scores would not be indicative of their relative physics knowledge based on written work and oral explanations. What our results do show is, that for those students who were involved in the validity study, their scores and relative rankings from the multiple-choice test were consistent with the scores from having their work graded and having the students interviewed.

IV. DISCUSSION

The reliability and validity studies verify that the multiple-choice exams administered in the introductory physics courses at the University of Illinois are fulfilling their primary function of assessing student understanding and assigning the appropriate grade. One should be careful not to conclude from these results that seeing and grading student work is not important. In addition to changing the exams to multiple-choice format, the course reform included the transformation of the recitations sections into discussion sections.^{27,28} The discussion sections have students working in groups of four on concepts and calculations. The emphasis of these sections is on showing work and justifying reasoning and strategies. Students receive feedback on this work from classmates as well as the teaching assistant. Each discussion section ends with a constructed-response quiz, which is graded by the teaching assistant based on the work shown.

It might appear that the reform just shifted the grading difficulty from exams onto quizzes. Certainly grades for the constructed response quizzes suffer the same reliability shortcoming as the exams. Perhaps even more, as a single TA grades all of the quizzes for an individual student

throughout the semester. However, the impact of the quizzes on the final grade is significantly less than the exams. We see the role of the quiz as more of a formative rather than an evaluative assessment. In addition, since quizzes are given every week, the grade on any individual problem has a very small impact on the student's course grade. The result is that both students and faculty are generally satisfied with the quiz format and grading.

V. CONCLUSIONS

This study demonstrates that physics instructors under real time constraints can produce multiple-choice exams which yield results that are both reliable and valid assessments of students' understanding of introductory physics. Statistics such as the Cronbach α provide a straightforward method for determining the reliability of exam scores, and hence the statistical uncertainty in any student's score. Integrating all questions over the course of a semester reveals that students' total exam score uncertainty is about 3%, which corresponds to a course grade uncertainty of roughly $\frac{1}{4}$ of a full letter grade.

Assessing the validity of exam scores is much more difficult as it requires making an independent assessment of the student's physics knowledge with which to compare the exam results. Although this is not practical to do, in general, a study of 33 students taking the calculus-based E&M course at the University of Illinois, who had scored consistently on their three midterm exams, showed that the multiple-choice exams gave a statistically equivalent assessment of their understanding compared to their written explanations and interviews. Indeed, the difference between these rankings was less than the statistical difference of 3% found in the reliability analysis. Although some "poor" questions inevitably make their way into exams, the large number of questions throughout the course provides sufficient information to accurately assess students' understanding.

ACKNOWLEDGMENTS

The authors would like to thank Jose Mestre, Eugene Torigoe, Adam Feil, and Eric Potter for useful discussions.

APPENDIX A: TEST CONSTRUCTION AND EVALUATION

Although multiple-choice exams are easier to grade than free-response exams, they are also more difficult to create. On a constructed-response exam, poorly worded or easily misinterpreted questions can be compensated for during the grading. Multiple-choice exams do not have this flexibility. Hence, the preparation of good multiple-choice questions is essential to the reliability and validity of the exams. The team-teaching approach of the introductory physics courses helps ensure this quality.

The introductory physics courses at UIUC are taught by a team of three to four professors, depending on class size. One or two professors are responsible for lecturing, one professor is in charge of the laboratory teaching assistants, and

TABLE III. A list of the types of questions used on midterm exams in our calculus-based E and M course (1003 questions in total).

No. of choices	Type of question	Exhaustive	% Used
2	Qualitative	Y	22.5
3	Graphical	N	2.5
3	Qualitative	N	7.3
3	Qualitative	Y	15.0
3	Quantitative	N	4.4
3	Scaling	N	1.7
3	Symbolic	N	1.6
5	Graphical	N	1.7
5	Qualitative	N	2.1
5	Qualitative	Y	0.2
5	Quantitative	N	37.9
5	Scaling	N	0.4
5	Symbolic	N	2.8

another is in charge of the discussion teaching assistants. In addition to their other assignments, this team is also responsible for creating the exams. It should be noted that in the four introductory courses between the years of 1999 and 2003, more than 50 physics professors contributed in creating the exams. Of these professors, most know very little of the research that has been done in the creation of questions with good distractors.

Each professor is typically assigned a few topics on which to write problems. They are encouraged to define an interesting situation, and then ask several questions that pertain to the situation. These problems are then assembled into an exam, which is reviewed by each of the team members. Having several independent people review the exam typically results in significant improvements to the questions.

The types of questions that appear on the exams are qualitative, quantitative, graphical, symbolic, and scaling questions with two-, three-, or five-choice answers to choose from.²⁹ Sometimes these answers exhaust all possible answers and sometimes they do not. Some examples of questions used on the exams can be found in Appendix B.³⁰ Table III is a listing of the various types of questions that have appeared on midterm exams given in the calculus-based E&M course from the spring of 1997 through the fall of 2002. In the table, the questions are listed by their number of choices, their type (qualitative, quantitative, etc.), whether the choices exhaust all possible answers, and what percentage these types of questions have appeared on the various exams.

After the exam has been administered, a standard exam analysis is performed and made available to the professors. In addition to the average exam score and average question score, the Cronbach α is provided as well as a discrimination analysis for each question. Figure 5 shows a typical discrimination analysis. The class is broken up into groups of 50 students based on their exam score. Each group's average score on that question is plotted versus their average exam score. Questions with good discrimination have a steep

FIG. 5. (Color) This is a discrimination plot for question 23 on the first final exam shown in the Appendix. Each data point represents a bin of approximately 50 students. The exam score for each bin is the average for that bin on the remaining questions on the exam.

slope, questions with little discrimination are relatively flat and often deserve a second look. Sometimes questions with low discrimination are simply “unique.” For instance, a fact about one of the laboratories might have low discrimination. Sometimes, however, they reveal an ambiguous or misleading question. This is important feedback which helps improve future exams.

TABLE IV. A list of each student’s multiple-choice (MC), average-grader (AG), and committee (CS) score for those who participated in the validity study. The data for each group are shown separately.

Group 1				Group 2			
Student	MC	AG	CS	Student	MC	AG	CS
1	59	55	75	1	66	58	58
2	92	95	99	2	73	71	65
3	71	65	70	3	100	88	95
4	81	70	70	4	84	82	84
5	65	57	57	5	29	38	47
6	100	97	96	6	69	58	52
7	68	54	52	7	73	71	78
8	58	60	67	8	53	53	61
9	76	87	91	9	55	50	51
10	57	41	38	10	100	95	95
11	70	64	76	11	62	66	67
12	65	47	53	12	56	53	52
13	87	72	91	13	51	54	71
14	63	41	44	14	53	50	55
15	54	46	55	15	69	56	70
16	76	79	98	16	77	81	83
				17	68	50	55

FIG. 6. (Color) The raw data from the students involved in the validity study ($N=33$). The graph is a plot of the students' average-grader score (AG) versus their multiple-choice score (MC).

To conclude, we offer a conjecture as to why these multiple-choice exams that contain questions that were constructed without the use of research-based distractors, can nonetheless, be valid and reliable. The first point to make is that about 65% of the nonquantitative questions have choices that exhaust all possible answers. Clearly, the issue of research-based distractors is moot for these questions. Indeed, we see that the number of these exhaustive questions that have poor discrimination is about 50% less than that from the nonexhaustive ones. The second point to make is that instructors are encouraged to construct qualitative questions with answers, which if nonexhaustive, at least are couched in general terms and avoid specific explanations.

APPENDIX B: VALIDITY STUDY DATA

In Table IV we provide the raw data from our validity study. The table lists each students' multiple-choice, average-grader, and committee score. Plots of this data can be seen in Figs. 6 and 7.

APPENDIX C: FINAL EXAMS QUESTIONS USED IN THE VALIDITY STUDY

1. Final 1

The following 20 test items are a subset of questions taken from the first version of the spring 1999 semester final exam in the calculus-based E&M course. These items were used in the validity study.

3. If the magnetic flux through a coil is zero at time t_0 , the induced current in the coil must also be zero at time t_0 .
(A) True (B) False

FIG. 7. (Color) The raw data from the students involved in the validity study ($N=33$). The graph is a plot of the students' committee score (CS) versus their multiple-choice score (MC).

The next two questions pertain to the following situation:

Three identical rectangular wire loops ($b > a$) are being moved in the plane of the page at speed v into a $\mathbf{B}$ field filled (shaded) region from a region of zero $\mathbf{B}$ field. The $\mathbf{B}$ field in the shaded region is spatially uniform and is normal to and pointing out of the plane of the page. When each loop is exactly half way into the shaded region:

5. The direction (clockwise or counterclockwise) of the current being induced in loop 2 is the same as the direction of the current being induced in loop 1.

(A) True (B) False

6. The magnitude of the current being induced in loop 3 is greater than the magnitude of the current being induced in loop 1.

(A) True (B) False

10. W is the network you would have to do to move the charges from configuration I, to configuration II.

Which one of the following is true?

- (A) $W < 0$
(B) $W = 0$
(C) $W > 0$

15. Two parallel conducting rails are connected by a resistor R in a region where there is a constant and spatially uniform B field perpendicular to and directed into the plane of the rails, as shown.

When a conducting bar in electrical contact with the rails is being pushed toward the resistor:

- (A) there is current flowing from **a** through R to **b**
- (B) there is no current in R
- (C) there is current flowing from **b** through R to **a**

16. A copper ring is being rotated clockwise by an external agent at a constant angular speed around a point on the ring as shown in the figure below. The ring is in the plane of the page and its motion is also in the plane of the page. The region is filled with a spatially uniform magnetic field normal to and pointing out of the plane of the page.

The current induced in the ring is

- (A) in the clockwise direction
- (B) in the counterclockwise direction
- (C) zero

17. The circuit shown in the diagram lies fixed in the plane of the page except for the semicircle of wire in the top side of the circuit which can be rotated around the axis defined by the top side of the circuit by the crank shown at the right. All this is in a region of space completely filled with a spatially uniform magnetic field normal to and pointing out of the plane of the page.

When the crank is turned by an external agent at a constant angular frequency ω , what current flows in the resistor?

- (A) a current that is constant as long as ω is constant
- (B) a sinusoidally varying current of angular frequency ω
- (C) no current

The next three questions pertain to the figure below:

A positive charge of magnitude q is placed at (x,y)

$= (a,0)$ and a negative charge of magnitude $2q$ is placed at $(x,y) = (-a,0)$ as shown in the figure above. The numerical values are $q = 3 \mu\text{C}$, $a = 5 \text{ cm}$.

18. There will be no place on the x axis for $-a < x < +a$ at which the **net electric field** due to these charges is zero.

- (A) True
- (B) False

19. There will be no place on the x axis for $x > +a$ at which the **net electric field** due to these charges is zero.

- (A) True
- (B) False

20. What is the value of E_y , the y component of the electric field due to these two charges at point A defined as $(x,y) = (0,-2a)$? Be careful—all the answers can be attained using values given in the problem.

- (A) $E_y = +5.79 \times 10^6 \text{ N/C}$
- (B) $E_y = +1.93 \times 10^6 \text{ N/C}$
- (C) $E_y = 0 \text{ N/C}$
- (D) $E_y = -1.93 \times 10^6 \text{ N/C}$
- (E) $E_y = -5.79 \times 10^6 \text{ N/C}$

The next four questions pertain to the following situation:

A solid metal sphere of radius a has a net **positive** charge Q_a . The sphere is surrounded by a thin concentric conducting spherical shell of radius b . The shell has a net **negative** charge $Q_b = -Q_a$.

21. Various spherical Gaussian surfaces are drawn concentric to the conducting sphere and shell at different radii R . Which graph best describes the electric **flux** Φ through the **entire** Gaussian surface as a function of R ? (Recall that the area vector for a closed Gaussian surface points outward.)

22. Let $a = 2 \text{ cm}$, $b = 5 \text{ cm}$, $Q_a = +2 \times 10^{-9} \text{ C}$, and $Q_b = -2 \times 10^{-9} \text{ C}$. Calculate the radial component of the **electric field** at $R = 4 \text{ cm}$ due to the conducting shell and sphere.

- (A) $E_r = -2.91 \times 10^4 \text{ N/C}$
 (B) $E_r = -1.55 \times 10^4 \text{ N/C}$
 (C) $E_r = 0$
 (D) $E_r = +1.12 \times 10^4 \text{ N/C}$
 (E) $E_r = +2.35 \times 10^4 \text{ N/C}$

23. Calculate the **electric potential** at the origin, given that the potential at infinity is zero.

- (A) $V = +540 \text{ V}$
 (B) $V = +267 \text{ V}$
 (C) $V = 0$
 (D) $V = -108 \text{ V}$
 (E) $V = -405 \text{ V}$

24. A spherical Gaussian surface (shown with the dotted line in the figure) is drawn concentric to the conducting sphere and shell at a radius $R = 4 \text{ cm}$.

When a positive point charge $+Q_a$ is brought close to (but outside of) the conducting shell the magnitude of the electric **flux** through the **entire** Gaussian surface at $R = 4 \text{ cm}$

- (A) increases
 (B) decreases
 (C) remains the same

The next three questions pertain to the following situation:

25. Calculate $|U_I|$ the magnitude of the potential energy for the configuration of charges shown as (I), given $Q = 3 \mu\text{C}$ and $d = 2 \text{ meters}$.

- (A) $20.0 \times 10^{-3} \text{ J}$
 (B) $28.6 \times 10^{-3} \text{ J}$
 (C) $37.3 \times 10^{-3} \text{ J}$
 (D) $46.0 \times 10^{-3} \text{ J}$
 (E) $63.4 \times 10^{-3} \text{ J}$

26. The potential energy of the configuration of charges shown as (II) is

- (A) $U_{II} > 0$
 (B) $U_{II} = 0$
 (C) $U_{II} < 0$

27. Compare U_I the potential energy of configuration I with U_{II} the potential energy of configuration II.

- (A) $U_I > U_{II}$
 (B) $U_I = U_{II}$
 (C) $U_I < U_{II}$

28. Shown below is a portion of a very thin infinite charged insulating sheet perpendicular to the x axis. The sheet has uniform positive charge density $+\sigma_a = +10 \mu\text{C/m}^2$. An infinite conducting slab, with thickness 1 cm , is also placed perpendicular to the x axis and 10 cm to the left of the insulating sheet as shown in the figure. The total surface charge density on the conducting slab, $\sigma_L + \sigma_R$, is $-6 \mu\text{C/m}^2$.

What is the surface charge on only the right side of the conducting slab (the side closest to the sheet)? Be careful— all the answers can be attained using values given in the problem.

- (A) $\sigma_R = 0$
 (B) $\sigma_R = -3.0 \mu\text{C/m}^2$
 (C) $\sigma_R = -6.0 \mu\text{C/m}^2$
 (D) $\sigma_R = -8.0 \mu\text{C/m}^2$
 (E) $\sigma_R = -10.0 \mu\text{C/m}^2$

37) An ac generator consists of a square coil with $N = 25$ turns and side dimension $b = 3 \text{ cm}$ in a spatially uniform magnetic field $B = 0.45 \text{ T}$ that points in the positive z direction. The coil rotates about the x axis at constant angular frequency $\omega = 666 \text{ rad/s}$.

Calculate the magnitude of the peak **EMF** generated in the coil.

- (A) $EMF_0 = 4.71 \text{ V}$
 (B) $EMF_0 = 5.36 \text{ V}$
 (C) $EMF_0 = 6.74 \text{ V}$
 (D) $EMF_0 = 7.82 \text{ V}$
 (E) $EMF_0 = 9.45 \text{ V}$

41. Suppose we have two point charges located along the y axis: Q_A at $y = +a$ and Q_B at $y = -a$. Which of the following statements about the signs and magnitudes of the charges must be true if the electric field produced by these two charges is equal to zero at $(x, y) = (0, +2a)$?

- (A) Q_A and Q_B have the *same sign* and the magnitude of Q_A is *less than* the magnitude of Q_B .
 (B) Q_A and Q_B have the *same sign* and the magnitude of Q_A is *greater than* the magnitude of Q_B .
 (C) Q_A and Q_B have the *opposite sign* and the magnitude of Q_A is *less than* the magnitude of Q_B .
 (D) Q_A and Q_B have the *opposite sign* and the magnitude of Q_A is *greater than* the magnitude of Q_B .

There is not enough information given to determine both the relative signs and the relative magnitudes of Q_A and Q_B .

2. Final 2

The following 20 test items are a subset of questions taken from the second version of the spring 1999 semester final exam in the calculus-based E&M course. These items were used in the validity study.

3. A wire coil is located in an external magnetic field. If the magnetic flux through this coil is zero at time t_0 , the induced current in the coil must also be zero at time t_0 .

- (A) True (B) False

4. Three identical copper loops are leaving a region of uniform magnetic field at the instant shown. The loops all have the **same speed**. Assume the magnetic field is uniform inside the region and zero outside.

The induced current is clockwise in all three loops.

- (A) True (B) False

8. W is the network you would have to do to move the charges from configuration I, to configuration II.

Which one of the following is true?

- (A) $W < 0$
(B) $W = 0$
(C) $W > 0$

The next 3 problems pertain to the situation below:

Consider two isolated well separated (i.e., neglect any effect of one sphere on the other) solid spheres of equal radii R each carrying total positive charge Q . One sphere is conducting, the other sphere is insulating (with the charge distributed uniformly throughout the volume).

9. If the potential at $r = \infty$ is zero, which of the following statements is true about the potential outside the radius of the sphere (i.e., for all $r > R$)?

- (A) $V(r > R)_{\text{conducting}} < V(r > R)_{\text{insulating}}$
(B) $V(r > R)_{\text{conducting}} = V(r > R)_{\text{insulating}}$
(C) $V(r > R)_{\text{conducting}} > V(r > R)_{\text{insulating}}$

10. If the potential at $r = \infty$ is zero, which of the following statements is true about the potential at the center of the spheres?

- (A) $V(r=0)_{\text{conducting}} < V(r=0)_{\text{insulating}}$
(B) $V(r=0)_{\text{conducting}} = V(r=0)_{\text{insulating}}$
(C) $V(r=0)_{\text{conducting}} > V(r=0)_{\text{insulating}}$

11. If the potential at $r = \infty$ is zero, what is the potential inside the conducting sphere?

- (A) $V(r < R)_{\text{conducting}} < 0$
(B) $V(r < R)_{\text{conducting}} = 0$
(C) $V(r < R)_{\text{conducting}} > 0$

14. Two parallel conducting rails in a horizontal plane are connected by a resistor R . They are in a region of spatially-uniform magnetic field that points out of the page as shown in the figure.

A conducting bar in electrical contact with the rails is being pulled away from the resistor at constant speed v by an external agent. Which one of the following is true?

- (A) A current flows through R in the direction of arrow 1.
(B) A current flows through R in the direction of arrow 2.
(C) No current flows through R as long as v is constant.

16. A copper ring is being rotated clockwise by an external agent at a constant angular speed around a point on the ring as shown in the figure below. The ring is in the plane of the page and its motion is also in the plane of the page. The region is filled with a spatially uniform magnetic field normal to and pointing out of the plane of the page.

The current induced in the ring is

- (A) in the clockwise direction
(B) in the counterclockwise direction
(C) zero

17. The circuit shown in the diagram lies fixed in the plane of the page except for the semicircle of wire in the top side of the circuit which can be rotated around the axis defined by the top side of the circuit by the crank shown at the right. All this is in a region of space completely filled with a spatially uniform magnetic field normal to and pointing out of the plane of the page.

When the crank is turned by an external agent at a constant angular frequency ω , what current flows in the resistor?

- (A) a current that is constant as long as ω is constant
 (B) a sinusoidally varying current of angular frequency ω
 (C) no current

The next three questions pertain to the figure below:

A positive charge of magnitude $2q$ is placed at $(x, y) = (a, 0)$ and a negative charge of magnitude q is placed at $(x, y) = (-a, 0)$ as shown in the figure above. The numerical values are $q = 3 \mu\text{C}$, $a = 5 \text{ cm}$.

20. There will be at least one place on the x axis for $-a < x < +a$ at which the **net electric field** due to these charges is zero.

- (A) True (B) False

21. There will be at least one place on the x axis for $x > +a$ at which the **net electric field** due to these charges is zero.

- (A) True (B) False

22. What is the value of E_y , the y component of the electric field due to these two charges at point A defined as $(x, y) = (0, -2a)$? Be careful—all the answers can be attained by using values given in the problem.

- (A) $E_y = +5.79 \times 10^6 \text{ N/C}$
 (B) $E_y = +1.93 \times 10^6 \text{ N/C}$
 (C) $E_y = 0 \text{ N/C}$
 (D) $E_y = -1.93 \times 10^6 \text{ N/C}$
 (E) $E_y = -5.79 \times 10^6 \text{ N/C}$

23. Shown below is a portion of a very thin infinite charged insulating sheet perpendicular to the x axis. The sheet has uniform positive charge density $+\sigma_a$. A cylindrical Gaussian surface (centered on the x axis) of length $2L_0$ encloses a portion of the sheet. The radius of each end cap is R . (Recall that for a closed surface, the area vector points outward.)

The electric flux through the left end cap (surface 2) of the Gaussian surface is

- (A) positive
 (B) negative
 (C) zero

24. A positive point charge Q_0 is now placed 10 cm to the right of the sheet and on the x axis as shown in the figure below (ignore the **X** in the figure until the next problem).

Assume the charge distribution on the sheet is unaffected by the point charge $+Q_0$. The absolute value of the flux through the **left end cap** (surface 2) will

- (A) increase
 (B) decrease
 (C) remain the same

25. Let $+Q_0 = +2 \mu\text{C}$ ($+\sigma_0$ is still $+5 \mu\text{C/m}^2$). What is the magnitude of the **net** electric field on the x axis a distance 10 cm to the **left** of the plane (at the **X** in the figure)? Be careful—all the answers can be attained using values given in the problem.

- (A) $E = 1.68 \times 10^5 \text{ N/C}$
 (B) $E = 2.82 \times 10^5 \text{ N/C}$
 (C) $E = 4.50 \times 10^5 \text{ N/C}$
 (D) $E = 7.32 \times 10^5 \text{ N/C}$
 (E) $E = 9.78 \times 10^5 \text{ N/C}$

The next two questions pertain to the figure below:

A thin conducting spherical shell of radius $a = 3 \text{ cm}$ has a net charge $Q_a = +3 \times 10^{-9} \text{ C}$. The inner shell is surrounded by a thin concentric conducting spherical shell of radius $b = 7 \text{ cm}$. The outer shell has a net charge $Q_b = -3 \times 10^{-9} \text{ C}$.

26. Calculate the radial component of the electric field at $r = 5 \text{ cm}$.

- (A) $E_r = +2.26 \times 10^4 \text{ V/m}$
 (B) $E_r = +1.08 \times 10^4 \text{ V/m}$
 (C) $E_r = 0$
 (D) $E_r = -1.49 \times 10^4 \text{ V/m}$
 (E) $E_r = -2.79 \times 10^4 \text{ V/m}$

27. Calculate the **electric potential** at the origin, given that the potential at infinity is zero.

- (A) $V = -386 \text{ V}$
 (B) $V = -103 \text{ V}$
 (C) $V = 0$
 (D) $V = +254 \text{ V}$
 (E) $V = +514 \text{ V}$

28. Shown below is a portion of a very thin infinite charged insulating sheet perpendicular to the x axis. The sheet has uniform positive charge density $+\sigma_a = +5 \mu\text{C}/\text{m}^2$. An infinite conducting slab, with thickness 1 cm, is also placed perpendicular to the x axis and 10 cm to the left of the insulating sheet as shown in the figure. The total surface charge density on the conducting slab, $\sigma_L + \sigma_R$, is $-3 \mu\text{C}/\text{m}^2$.

What is the surface charge on only the right side of the conducting slab (the side closest to the sheet)? Be careful— all the answers can be attained using values given in the problem.

- (A) $\sigma_R = 0$
- (B) $\sigma_R = -1.5 \mu\text{C}/\text{m}^2$
- (C) $\sigma_R = -3.0 \mu\text{C}/\text{m}^2$
- (D) $\sigma_R = -4.0 \mu\text{C}/\text{m}^2$
- (E) $\sigma_R = -5.0 \mu\text{C}/\text{m}^2$

33. Suppose we have two point charges located along the y axis: Q_A at $y = +a$ and Q_B at $y = -a$. Which of the following statements about the signs and magnitudes of the charges must be true if the electric field produced by these two charges is equal to zero at $(x, y) = (0, -2a)$?

- (A) Q_A and Q_B have the *same sign* and the magnitude of Q_A is *less than* the magnitude of Q_B .
- (B) Q_A and Q_B have the *same sign* and the magnitude of Q_A is *greater than* the magnitude of Q_B .
- (C) Q_A and Q_B have the *opposite sign* and the magnitude of Q_A is *less than* the magnitude of Q_B .
- (D) Q_A and Q_B have the *opposite sign* and the magnitude of Q_A is *greater than* the magnitude of Q_B .
- (E) There is not enough information given to determine both the relative signs and the relative magnitudes of Q_A and Q_B .

37. An ac generator consists of a circular coil with $N = 30$ turns radius $R = 2$ cm in a spatially uniform magnetic field $B = 0.55$ T that points in the positive y direction. The coil rotates about the x axis at constant angular frequency $\omega = 333$ rad/s.

Calculate the **magnitude of the peak EMF** generated in the coil.

- (A) $\text{EMF}_0 = 4.82$ V
- (B) $\text{EMF}_0 = 5.49$ V
- (C) $\text{EMF}_0 = 6.90$ V
- (D) $\text{EMF}_0 = 8.01$ V
- (E) $\text{EMF}_0 = 9.68$ V

¹D. K. Campbell, C. M. Elliot, and G. E. Gladding, *Parallel parking an aircraft carrier: Revising the calculus-based introductory physics sequence at Illinois* (Forum on Education of the American Physical Society, 1997).

²R. Lukhele, D. Thissen, and H. Wainer, On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests, *J. Educ. Meas.* **31**, 234 (1994).

³E. F. Redish, *Teaching Physics with the Physics Suite* (John Wiley and Sons, New York, 2003).

⁴S. Tobias and J. B. Raphael, In-class examinations in college-level science: New theory, new practice, *J. Sci. Educ. Technol.* **5**, 311 (1996).

⁵G. J. Aubrecht and J. D. Aubrecht, Constructing objective tests, *Am. J. Phys.* **51**, 613 (1983).

⁶Midterm exams are written to be 60 min exams, but students are allotted 90 min to complete them. Students are allotted 3 h to take the final exam. For most students, time is not an issue.

⁷For clarification, to get each student's even and odd scores, each of the four exams were first ordered by item difficulty. Then a student's even score is the sum of their scores from the even questions from exams 1 and 3 and the odd questions from exams 2 and 4. Likewise, a student's odd score is the sum of their

scores from the odd questions from exams 1 and 3 and the even questions from exams 2 and 4.

⁸This analysis considers only our A to C students because it is these students whose exam performance shows a strong linear correspondence to their assigned letter grade. That is, these students tend to receive 90% or more credit on the effort components of the course (e.g., homework, quizzes, and laboratories). Thus, their effort grade is not a distinguishing factor to the grade they receive in the course. This is not true, in general, for D and F students. Not only do these students do poorly on the exams, they also tend to do poorly on the effort components of the class. Therefore, the strong linear relationship between exam performance and assigned letter grade that is present for A to C students is not present for D to F students.

⁹It should also be noted that over this same time span, more than 50 physics professors contributed in creating the exams used in the introductory courses.

¹⁰In a second splitting method, the "even" test is literally the collection of the even-numbered questions from the first and third midterms and the odd-numbered questions from the second midterm and final. The reverse construction is made for the "odd" test. The uncertainty found using this method was 3.5%.

- ¹¹A third splitting method is simply an alteration of the second splitting method. Here, the “even” test is questions 1, 4, 5, 8, 9,..., from the first and third midterms and questions 2, 3, 6, 7,..., from the second midterm and final. The reverse construction is made for the odd test. The uncertainty from this splitting was 3.6%.
- ¹²An offset to zero for each semester could be made so that all semesters had the same average percent difference in even and odd tests. This correction would account for the fact that students in different course semesters do not have the same even and odd tests. Adding this offset has the inherent effect of diminishing the standard deviations in the distributions to 3.2% for both the second and third methods of splitting the questions. This offset had little effect on the first splitting method.
- ¹³J. R. Taylor, *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements* (University Science Books, Sausalito, CA, 1982).
- ¹⁴A letter grade difference of 1.0 is equivalent to a letter grade difference of A to B or B to C. A letter grade difference of $\frac{1}{3}$ is equivalent to the difference between an A and an A– or an A– to a B+.
- ¹⁵H. Wainer and D. Thissen, in *True Score Theory: The Traditional Method*, edited by David Thissen and Howard Wainer (Lawrence Erlbaum Associates, Hillsdale, NJ, 2001), Chap. 2, pp. 23–72.
- ¹⁶C. C. Peters and W. R. Van Voorhis, *Statistical Procedures and Their Mathematical Bases* (McGraw-Hill, New York, 1940).
- ¹⁷Common convention is to desire reliability correlation coefficients greater than 0.80 to ensure that a student’s exam score uncertainty is less than half of the standard deviation in the class’ exam score distribution.
- ¹⁸L. J. Cronbach, Coefficient alpha and the internal structure of tests, *Psychometrika* **16**, 297 (1951).
- ¹⁹Because some of the exam items are grouped together under the same physical situation, splitting these items into separate split-half exams generally increases the correlation coefficient between the split-half exams and thus artificially increases the coefficient alpha. It may be more appropriate to treat those questions that are grouped together under the same prompt as testlets, and then to calculate alpha using testlet scores. To see what effect this might have on our alpha values, we examined four semester sets of exams: two from calculus-based mechanics and two from algebra-based mechanics. In each of the four semesters, the testlet alpha was indeed less than the item alpha, but never by more than 2% of the item alpha. This difference between the item and testlet alphas is less than the variation between semester item alphas.
- ²⁰T. P. Hogan, *Relationship between free-response and choice-type tests of achievement: A review of the literature* (ERIC Clearinghouse on Tests and Measurements, Princeton, NJ, 1981).
- ²¹One justification for this selection process is that if only A and F students participated in the study, correlations between multiple-choice and constructed-response scores would artificially be high. We wanted to make sure there was an even distribution of students in the letter grade range from A to C. This is the range of most interest to us since it is this range students’ course grades are predominately dependent upon exam performance. Students in the D to F range do poorly on all components of the course, not just the exams. To ensure that there were equal number of students in each grade category, we chose to select only those students who had scored consistently on their three midterm exams. If a student receives an “A” on one midterm but then receives a “C” on another, one does not know whether this student is really an A, B, or C student.
- ²²This weighting system was instituted to allow for partial credit. The five-option items are intended to be more difficult than two- and three-option items. Students can receive partial credit on a five-option item in one of the following ways: six points if only one option is chosen and is correct, three points if only two options are chosen and one of the chosen options is correct, two points if only three options are chosen and one of the chosen options is correct, and zero points for all other markings.
- ²³To address any concerns that these raw correlations are large because of the selection of students who participated in the study, there is a correction that can be made to estimate what the raw correlations would be if the students were a pure random sampling of the entire class. This correction of heterogeneity had little effect on our raw correlations: for group 1, $r=0.88$ went to 0.90, and for group 2, $r=0.92$ went to 0.89. We were able to test the validity of this correction from our reliability data and found that it predicted on average at most a value that was only $0.62\% \pm 0.07\%$ over the actual value.
- ²⁴*Educational Measurement*, edited by R. L. Thorndike (American Council on Education, Washington, D.C., 1971).
- ²⁵Using the heterogeneity correction, the raw correlation values between MC and CS went from 0.78 and 0.83 to 0.81 and 0.77 for groups 1 and 2, respectively.
- ²⁶S. Eidelman, K. G. Hayes, K. A. Olive, M. Aguilar-Benitez, C. Amsler, D. Asner, K. S. Babu, R. M. Barnett, J. Beringer, P. R. Burchat, C. D. Carone, C. Caso, G. Conforto, O. Dahl, G. D’Ambrosio, M. Doser, J. L. Feng, T. Gherghetta, L. Gibbons, M. Goodman, C. Grab, D. E. Groom, A. Gurtu, K. Hagiwara, J. J. Hernández-Rey, K. Hikasa, K. Honscheid, H. Jawahery, C. Kolda, Y. Kwon, M. L. Mangano, A. V. Manohar, J. March-Russell, A. Masoni, R. Miquel, K. Mönig, H. Murayama, K. Nakamura, S. Navas, L. Pape, C. Patrignani, A. Piepke, G. Raffelt, M. Roos, M. Tanabashi, J. Terning, N. A. Törnqvist, T. G. Trippe, P. Vogel, C. G. Wohl, R. L. Workman, W.-M. Yao, P. A. Zyla, B. Armstrong, P. S. Gee, G. Harper, K. S. Lugovsky, S. B. Lugovsky, V. S. Lugovsky, A. Rom, M. Artuso, E. Barberio, M. Battaglia, H. Bichsel, O. Biebel, P. Bloch, R. N. Cahn, D. Casper, A. Cattai, R. S. Chivukula, G. Cowan, T. Damour, K. Desler, M. A. Dobbs, M. Drees, A. Edwards, D. A. Edwards, V. D. Elvira, J. Erler, V. V. Ezhela, W. Fetscher, B. D. Fields, B. Foster, D. Froidevaux, M. Fukugita, T. K. Gaisser, L. Garren, H.-J. Gerber, G. Gerbier, F. J. Gilman, H. E. Haber, C. Hagmann, J. Hewett, I. Hinchliffe, C. J. Hogan, G. Höhler, P. Igo-Kemenes, J. D. Jackson, K. F. Johnson, D. Karlen, B. Kayser, D. Kirkby, S. R. Klein, K. Kleinknecht, I. G. Knowles, P. Kreitz, Yu. V. Kuyanov, O. Lahav, P. Langacker, A. Liddle, L. Littenberg, D. M. Manley, A. D. Martin, M. Narain, P. Nason, Y. Nir, J. A. Peacock, H. R. Quinn, S. Raby, B. N. Ratcliff, E. A. Razuvaev, B. Renk, G. Rolandi, M. T. Ronan, L. J. Rosenberg, C. T. Sachrajda, Y. Sakai, A. I. Sanda, S. Sarkar, M. Schmitt, O. Schneider, D. Scott, W. G. Seligman, M. H. Shaevitz, T. Sjöstrand, G. F. Smoot, S. Spanier, H. Spieler, N. J. C. Spooner, M. Srednicki, A. Stahl, T. Stanev, M. Suzuki, N. P. Tkachenko, G. H. Trilling, G. Valencia, K. van Bibber, M. G. Vincet, D. Ward, B. R. Webber, M. Whalley, L. Wolfenstein, J. Womersley, C. L. Woody, O. V. Zenin, and R.-Y. Zhu, Review of Particle

- Physics, Phys. Lett. B **592**, 1 (2004).
- ²⁷P. Heller and M. Hollabaugh, Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving, Am. J. Phys. **60**, 627 (1992).
- ²⁸P. Heller and M. Hollabaugh, Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups, Am. J. Phys. **60**, 637 (1992).
- ²⁹Full credit for a two-choice, three-choice, or five-choice question is two points, three points, or six points, respectively. See end-note in the subsection “The Study” of the Validity section for an explanation of the weighted grading system.
- ³⁰For more examples of questions used in our exams, visit the Illinois Physics Education Research Group’s website at <http://www.physics.uiuc.edu/Research/PER/> and click on the “Resources” link. Researchers and teachers can gain free access to all of the midterm exams used in the introductory courses in recent years.