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

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 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 r0.94 at the 95% confidence level and oral interviews r=0.94

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