Validated diagnostic test for introductory physics course placement

Eric Burkholder,[1] Karen Wang,[2] and Carl Wieman[1,2]

[1]*Department of Physics, Stanford University, Stanford, California 94305, USA*
[2]*Graduate School of Education, Stanford University, Stanford, California 94305, USA*

(Received 28 December 2020; accepted 2 April 2021; published 19 April 2021)

Diagnostic tests for placing students into appropriate courses are commonly used in higher education math and science departments. We found no physics education research examining the validity of such exams, how well they are placing students into appropriate courses, and what guidance these exams can give for better focusing instruction. In this study, we present criteria by which to evaluate the validity and value of a diagnostic test, and we then apply these to the test developed for physics course placement at Stanford University. We examine the data collected from the first year in which all students intending to enroll in an introductory physics course were required to take the test and received an email suggesting which course would best match their level of preparation based on their score. We found that most students followed this advice. We also found that the diagnostic test score was a good predictor of performance in both the introductory physics course for physics majors and the course targeting science and engineering majors. It was not a good predictor of performance in two other courses, likely due to lack of alignment between course objectives and diagnostic measures. We found many individual questions which were particularly discriminating in predicting performance in the various courses, suggesting specific topics on which instruction should focus. We found that 27 of the 38 questions were predictive of performance in at least one of the courses studied, and only 8 questions that were not predictive of course performance or did not show any difference between students enrolled in different courses. These results provide evidence for the validity and usefulness of our diagnostic test, and we offer it to instructors at other institutions who might find it of value.

DOI: 10.1103/PhysRevPhysEducRes.17.010127

I. INTRODUCTION

Diagnostic tests for the purpose of placement are widely used in higher education. They are used in math departments at most institutions, primarily but not entirely to determine if students have to take "remedial" math courses. Providing such math placement exams is a large commercial enterprise [1]. They are less frequently but commonly used in chemistry [2] to decide if students should take a remedial chemistry, or sometimes, prechemistry math courses. We also find references in the literature to their use in biology for determining which introductory courses students can skip over [3]. We can find almost nothing in the physics education literature on the use of diagnostic placement tests, but we are aware of a number of institutions that use such tests to advise students as to what introductory physics track they should take, and/or what math preparation they need, in preparation for introductory physics [4]. The use of such diagnostic tests raises several interesting questions that have not been well studied. How good are they at placing students in the course that is best suited to their preparation? How valid are they and what should be the criteria to use to establish their validity? What can one learn from them to guide improvements in instruction? What areas of math and physics preparation are the most important to student success in introductory physics courses, and does this depend on the type of course? This paper addresses these questions.

The research literature on the validity of math placement exams is primarily about the effectiveness of commercial tests at predicting failure rates in precalculus courses, particularly at two-year colleges. They do not appear to be better than the ACT or SAT math exams and are relatively poor predictors of grades in math courses, with values of R squared (fraction of the variance they explain) in the 0.1 to 0.2 range [1,4–6]. Most Ph.D. granting institutions have "homemade" math placement tests, and we can find no research on their validity. In most studies of math test validity, there is the complication that the mandatory assignment of courses based on math diagnostic test scores is so prevalent. This imposes rigid selection effects on the populations being analyzed. As best we can determine, in other fields placement exams are more often used to advise students, but not completely determine, which courses they can and cannot take. There is a modest chemistry research literature on the validity of diagnostic tests. It is almost entirely limited to looking at the correlation between placement test scores and either course grades or, more commonly, failure rates in general chemistry [2,7]. We can find no physics education research looking at the validity of diagnostic placement exams in physics.

We recognize that there are varying opinions as to whether it is appropriate to have different streams of introductory math and science courses into which students are sorted. The placing of students into "remedial math" is seen as particularly problematic, with some indications that it does more harm than good [4]. However, most large physics departments have some combination of the following introductory course sequences, "physics for nonscientists," "algebra-based physics," often primarily intended for premed students, "physics for science and engineering students," and "physics for aspiring physics majors or honors."

Usually, departments will also have some criteria for advising students as to the course that is best suited to their preparation. Personally, we are somewhat uncomfortable with the practice of having different tracks and a test for placing students, but we are not addressing such concerns in this paper. We simply accept that many institutions, including ours, have and will continue to have such practices. Given that context, it is important to determine if such placement tests are achieving their intended purpose, and how might they also provide formative assessment on the instruction? We are assuming the intended purpose is placing students into a course in which, given their preparation, they will find challenging, and thus of substantial educational value, but not so challenging that they will not have a reasonably high probability of success.

Stanford University has several different tracks of introductory physics courses and has students take a diagnostic placement test before registering for any of these courses. The test covers a range of introductory physics topics and the math used in these introductory physics courses. Students are given recommendations, but not requirements, as to which course is most suitable for them, based on their exam score. The data provided by this exam and student performance in the different courses were used to address the questions posed above.

Our primary research questions were

1. How valid is the Stanford physics diagnostic test?
2. How well does it diagnose specific important areas of preparation?

The unique context and goals of a placement exam make it rather different from summative exams, and hence subject to different requirements for validity. First, the construct that it is to measure is not what does the student know, but rather, how well prepared are they to learn in a particular course? Second, it must measure this for a very wide range of student backgrounds, and it must measure their preparation to learn, not in a single course, but rather for each of several different courses. Third, it is essentially zero stakes and taken outside the context of any course. That means that the amount of effort and attention given to it can be quite variable, and hence can be a substantial factor on performance, which implies that the exam must be as short as possible, to maximize attention. Because these are administered to a very large number of students, it effectively means they must also be computer graded. While the commercial math placement exams use adaptive computer testing programs, in the physics context, multiple choice exams are the only practical option. As shown below, this does not appear to be a serious limitation, as the predictive value of the Stanford physics exam is substantially higher than reported for math placement tests.

These issues, in combination, imply that the exam must probe a very wide range of topics at a very wide range of levels of difficulty, using as few questions as possible. This guarantees that standard psychometric test of reliability and validity will not be very meaningful. [8] For example, for standard summative exams, it is desirable to have a high value for Cronbach alpha. This says that there is a good correlation between responses to different questions, and so they are measuring the same construct. For a diagnostic test, a high value for Cronbach alpha is undesirable. It means that there are too many questions covering similar things, so it should either be made shorter by dropping some questions or those questions should be replaced with some that probe other topics that will have less correlations and overlap in responses. Though it is possible to have a good test with a low alpha value, we note that Cronbach alpha for this particular test is excellent (0.91).

What are the criteria of validity for a placement exam? For the relevant courses, it should be predictive of performance of the population of students who are in each of the courses. This also means that it should correctly predict that students who take a course even though they scored below the recommended level on the diagnostic, should perform poorly. The evaluation of validity based on how well a test predicts performance on some quantitative measure other than the test itself is fundamentally different than the assumptions underlying most conventional psychometric tests of an assessment instrument. The standard psychometric analyses assume that one only has a single quantitative measure, the assessment instrument itself. Hence those analyses focus on various statistical tests that measure the internal consistency of the test. However, with a diagnostic test, there is an external standard against which the exam should be analyzed. How well it predicts performance in the various courses with the least amount of exam time or questions is the primary criteria of validity.

If the score on the diagnostic test is a good predictor of performance in all the relevant physics courses, that says it is measuring factors that matter. The better it is as a predictor, the more completely and accurately it is measuring all the various factors that matter. However, no test will be able to measure every factor that determines performance. So, another test of validity is to see how well the test is capturing as many relevant factors as is practical. Thus, the second test of validity of the exam is to gather whatever data is available on students that might in principle be included in a diagnostic and determine the extent that any of this additional data improves predictive power. If any of this additional data improves how well one can predict the student course performance, it means the diagnostic is missing something and so the diagnostic needs to be modified to capture the relevant information. For example, one would do a regression analysis adding to the diagnostic test scores the math SAT or ACT scores (hereafter simply "SAT scores") and conceptual inventory pre-scores of the students to see how well these expanded models improve the predictive power in a given course, compared to the diagnostic alone. Here we stress the "practical" aspect. For example, at many institutions high school GPA is a good predictor of academic performance, but we have found that it is not practical for the physics department to get students' high school GPAs so we do not consider it.

If the diagnostic test predicts a large fraction of the variation in student performance across the different courses, it means the exam is valid, but if it only explains a small fraction in a given course it does not necessarily mean that it is not valid. The exam could be accurately measuring all the preparation that matters, but how well that preparation determines the course performance will depend on how the course is taught, the extent to which the instructor is aware of relevant variations in student preparation, and how well he or she addresses those differences in their instruction. As noted below, the Physics 41E course was specifically designed to address weaknesses in physics preparation of the least-prepared students much more than was done in the instruction in the standard Physics 41 course. Thus, it was not surprising that the diagnostic test was a weaker predictor of performance in 41E.

In addition to using a diagnostic test for placement of students, it can in principle serve a quite different purpose, as formative assessment on the instruction. By identifying those questions for which student responses are a particularly strong predictor of performance in a given course, the test can indicate to the instructor those topics for which providing extra resources or additional instructional time would have the highest probability of improving student outcomes in the course. Correspondingly, looking at the predictive power of the diagnostic before and after such targeted interventions provides the instructor and/or department with a measure of the success of such interventions. A successful intervention would reduce the predictive power of the relevant diagnostic questions.

We illustrate this with a hypothetical example. The diagnostic shows that a student's score on the questions involving the vector dot product is an important predictor of their grade in Physics 1. In response, the instructor adds an extra homework assignment focusing on the vector dot product. The instructor looks at how well the dot product questions on the diagnostic predict final exam scores in the year before and the year after this homework was added. She sees that after this change the scores on the dot product diagnostic test questions are ¼ as correlated with final exam scores as they were in the year before, but the predictive power of the other diagnostic questions was the same. This indicates that the instructional change was effective.

Another way a good diagnostic test can provide formative feedback on instruction is by helping to keep the grading fair and consistent. We are aware of many introductory science, technology, engineering, and mathematics courses where the failure rates in the same course but in different years taught by different instructors had quite different failure rates. We have seen differences as large as a factor of 2 and involving nearly a quarter of the students in the class (the large differences are primarily in math and chemistry courses). The justification always given for the difference is that the students in the lower performing year were less prepared or just not as intelligent. In the absence of any data to the contrary, that explanation was accepted, rather than the more plausible explanation that differences were due to variations in the exams, grading standards, or quality of instruction. A good diagnostic test would provide data as to the extent of incoming variation in students' preparation, and thus offer grounds to argue against the unfairness of very different grade distributions in different years or different sections of the same course.

II. METHODS

A. Study context

The Stanford University physics department has multiple first year physics courses covering mechanics, including a physics for prephysics majors, a physics for pre-engineers and physical or computer science majors, and a physics for premeds (and a scattering of other majors). For the past two years, it also offered an additional course, a physics for pre-engineers with unusually weak preparation in physics taught by two of the coauthors (E. B. and C. W.). Roughly half of the Stanford undergraduate population will take one or more of these introductory mechanics physics courses. These students have a substantial spread in their math preparation and a very large spread in their physics preparation [9], ranging from no physics at all in high school to high grades in multiple AP physics courses. The department has used an optional diagnostic placement exam on an advisory basis for many years. For the 2019–2020 academic year covered by this study, all students were told they were required to take this diagnostic before they could register for any physics course, and they received an automatic letter based on their score, advising them as to which of the course options they appeared to be prepared for, but not mandating which they could enroll in. The scores corresponding to the different recommendations was somewhat arbitrary, based largely on what the instructors of the various courses in the past felt was essential that students know to be successful. To a large extent, students followed the recommendations. The diagnostic test was originally created based on collecting items that instructors of the various introductory courses thought were most relevant. Prior to the year studied here, two of us (E. B. and C. W.) updated the existing test. We kept many of the original questions, but deleted several that seemed redundant, and we added a number of new ones to cover a wider range of areas, specifically ones that our research had shown were particularly difficult for many of the less prepared Stanford students and some inspired by the physics education research literature. Our changes were reviewed and modified slightly by several other instructors familiar with the introductory courses. The diagnostic covers vector operations, the basics of differential and integral calculus, as well as Taylor series, and a range of physics topics which are traditionally covered in introductory mechanics courses. In addition, there are some questions on angular momentum, electrostatic forces, and magnetic fields which are there to identify which students are the best prepared and likely able to take the honors course (described below). The questions on Taylor series and vector cross and dot products are similarly included to identify students best prepared for the honors sequence.

In the 2019–2020 academic year, a total of 1522 students took the diagnostic test [10]. On average they spent 29 min on the exam, with a standard deviation of 22 min. We excluded the data from the 421 students who did not complete the diagnostic and the 127 students who spent less than 10 min on the test. 421 of the students who spent more than 10 min on the test enrolled in one of the four mechanics courses offered. An additional 259 students enrolled in one of these introductory courses but did not take the diagnostic test or spent less than 10 min on it. The total enrollment in introductory mechanics courses was 680. For all 680 of these students, we received student's course grades and SAT or ACT math scores from the Office of Institutional Research and Decision Support. Course grades were converted to a 4.3 scale, where a 4.3 was an A+, and each partial letter grade was 0.3 grade points lower. Students who took the course pass-fail were assigned a 3.0 if they passed, and a 1.0 if they did not pass—only 16 students elected to take any course pass or fail. There were six students who withdrew or took an incomplete, and they were assigned a grade of 2.0. We chose these placeholder values as we hypothesized they would be similar to students' actual grades had they completed the course for a letter grade. SAT and ACT scores were converted to national percentiles using freely available concordance tables [11]. For two of the courses, the Force and Motion Conceptual Evaluation (FMCE) [12] was also administered in the first week of class as a zero-stakes assessment. We provide a description of each mechanics course, the enrollment, and the available data in Table I.

Physics 61 is the honors mechanics and special relativity course for students who intend to major in physics or engineering physics. Students who enrolled in this course were expected to have a score of 5 on the AP Physics C—Mechanics exam, at least a score of 32 (out of 36) on the department diagnostic test, or have already taken another introductory mechanics course at Stanford. This course has a vector calculus corequisite. Topics covered include special relativity, forces, kinematics in 3D, dimensional analysis, momentum, energy and work, angular momentum and torque in 3D, and damped and force harmonic oscillators. The course uses the text by Kleppner and Kolenkow [13] as well as a supplement by Morin [14]. This course is taught using minilectures interspersed with active learning small group activities. This course has two midterms and one final which constitute 70% of the course grade. The remainder of the grade consists of weekly problem sets (25%) and participation (5%). Students in Phys 61 also enroll in the one-credit lab course, Phys 62. The lab course is required for physics majors, which comprise the majority of students enrolled in Phys 61.

Phys 41 is the standard calculus-based mechanics course for students intending to major in science and engineering, but it is also frequently taken by students intending to go to medical school. Students were advised that they should

TABLE I. Descriptions of all introductory mechanics courses offered at Stanford, their 2019-2020 enrollment, and how many students for whom we had (i) diagnostic scores, (ii) SAT or ACT math scores, and (iii) FMCE scores.

Course	Description	Enrollment	Diagnostic scores	SAT scores	FMCE prescores
Phys 61	Honors, calculus-based mechanics	68	52	68	Not applicable
Phys 41	Calculus-based mechanics for scientists and engineers	410	233	410	366
Phys 21	Algebra-based mechanics, thermodynamics, and fluids	106	67	106	Not applicable
Phys 41E	Calculus-based mechanics for students with minimal prior physics	96	69	96	73

receive at least 18 (out of 36) on the diagnostic test to enroll in this course. The math co-requisite for this course is Calculus 2, which covers integration. Phys 41 covers kinematics, forces and torques, momentum, angular momentum and uniform circular motion, and conservation of energy. The courses used the text by Young and Freedman [15]. The course is taught using primarily traditional lecture interspersed with clicker questions, with an additional weekly discussion section in which 15–18 students work on problems from Tutorials in Introductory Physics [16] guided by a TA. The course has two midterm exams and one final exam which constitute 80% of students' grades; the remainder of the grade is based on weekly problem sets and participation (~5%). This course was conducted in the winter of 2020, and thus the final week of instruction was remote and the final exam was canceled. The grading of other course components was kept in the same proportion; exam grades were the weighted averages of the first midterm (28%) and the second midterm (38%). Approximately 50% of the students in Phys 41 also enroll in Phys 42, the one-credit optional lab course that covers a portion of the Phys 41 material.

Phys 21 is an algebra-based course covering mechanics, heat, and fluids. The only requirements for this course are high school algebra and trigonometry. The course covers Newton's laws, uniform circular motion, projectile motion, static torques, momentum, work and energy, the first and second law of thermodynamics, ideals gases and kinetic theory, and the basics of hydrostatics and waves. The course uses the text by Giancoli [17]. The course is taught using active learning methods, which includes use of prelecture reading questions, and students completing worksheets in small groups during the class periods. The course has two midterm exams and one final exam, which constitute 77% of students' grades; the remainder of the grade is based on weekly problem sets and participation (~5%). Most students enrolled in Phys 21 also enroll in Phys 22, the separate one credit lab course covering the same material as Phys 21.

Phys 41E is a calculus-based course designed for students with no prior physics experience and less math background and is designed to replace Physics 41 for this student population. The course covers static forces and torques in depth, conservation of energy and work in depth, and 1D kinematics. The math corequisite is Calculus 2, as it is for Phys 41. This course uses no textbook, as the order of topics and their treatment is not mirrored in any existing texts but extensive supplementary materials were provided. This course is taught using cooperative group problem solving, with limited minilectures from the instructors. This course also incorporates learning assistants, who help during regular course sections and hold group and individual meetings with students outside of regular class time to provide additional support and help with study skills. This course has 4 in-class quizzes and a final exam, which together constitute 56% of the course grade. The remainder of the grade is 30% weekly homework and 15% participation, with up to 10% extra credit available. This course was conducted in the winter of 2020, and thus the final week of instruction was remote, and the final exam was canceled. The grading of other course components was kept in the same proportion; quiz grades were thus 38% of the total final grade. Some students also enroll in the Phys 42 optional lab course. There is no special lab course to accompany Phys 41E.

B. Analysis

RQ1: We used linear regression to test whether there was a linear relationship between diagnostic score and course outcomes (either course grade and/or exam grade, depending on the course). We quantified the strength of this relationship by R squared, which measures the proportion of variance in the outcome variable explained by the diagnostic score. We then used multiple linear regression to predict course outcomes as a function of diagnostic score, SAT score, and, if available, FMCE prescore. This allowed us to determine if other measures of incoming preparation were significant predictors of course performance after accounting for the diagnostic score. We determined if these other predictors are substantial by looking at the change in R squared between the first and second model. In all cases, we used multiple imputation to account for missing diagnostic or FMCE scores. We imputed 20 datasets using predictive mean matching and pooled the regression results across all 20 datasets. For more detail on multiple imputation, see Ref. [18].

RQ2: To answer our second research question, we computed the fraction of students in each course who got each question on the diagnostic correct—a measure which we call the "difficulty." We judged an individual question to be good at sorting students into the appropriate courses if there was a substantial difference between the difficulty measure for Phys 61, Phys 41, Phys 41E, and Phys 21. We also identified questions that were most useful in identifying students for Phys 61, as indicated by a difficulty measure for Phys 61 that is much higher than all the other courses. In addition to the difficulty parameter, we also calculated the "discrimination" of each test question for each course—the correlation between students' responses to that question and their overall course grades. If the discrimination was high (see Results below), we judged that item to be a good predictor of students' performance in that particular course.

III. RESULTS

In Fig. 1, we plot histograms of the total diagnostic score for each of the courses. We find that students in the honors course have the highest average scores (mean = 89%, s.d. = 6.0%), students in the standard calculus-based course

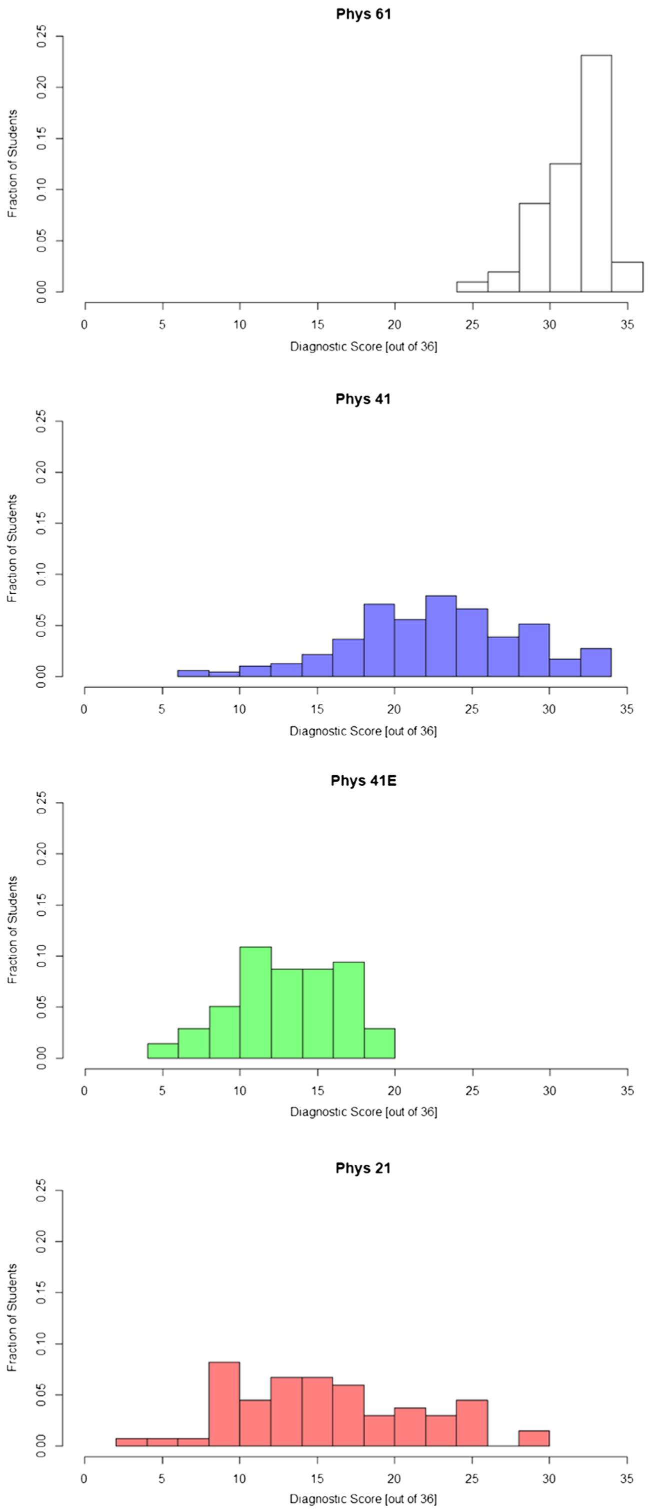

FIG. 1. Histogram of diagnostic scores for students in the honors course (Phys 61), standard calculus-based course (Phys 41), calculus-based course for poorly prepared students (Phys 41E), and algebra-based course (Phys 21).

have scores in the middle range of the instrument (mean = 63%, s.d. = 16%), and students in the algebra-based course (mean = 44%, s.d. = 17%) and calculus-based course designed for students with little to no prior physics (mean = 36%, s.d. = 9.4%) have the lowest scores. For comparison, the average FMCE score in Phys 41 is 36% (s.d. = 16%) and the average FMCE score in Phys 41E is 15% (s.d. = 7.3%).

C. Multiple linear regression

In Fig. 2, we plot students' course grades (for Phys 61 and 21) and exam grades (Phys 41 and 41E), converted to z scores, as a function of diagnostic scores. To guide the reader's eye, we also include trend lines calculated by ordinary least squares regression. With the exception of Phys 41E, it is clear that lower diagnostic scores are correlated with lower course performance (see trendlines). This points to the validity of the diagnostic for course placement. Students in Phys 41E have lower diagnostic scores than students in Phys 41. Because performance in Phys 41 strongly correlates with diagnostic score, one can thus conclude that Phys 41E students would perform poorly in Phys 41. Using similar logic, one can argue that Phys 41 students would perform poorly in Phys 61, except for the small fraction with unusually high diagnostic scores. We note that each trendline has a shaded gray region indicating the standard error on the prediction. Scores that fall within this region deviate from the average trend relating diagnostic score with course performance, but fall within a range suggesting that the trendline would reasonably predict that relationship.

The results of the multiple linear regression for Phys 61 are in Table II. For this course we only have course grades, but they are closely linked to exam scores, more so than in the other courses for the years we have data. We find that the diagnostic score is a strong predictor of course grade,

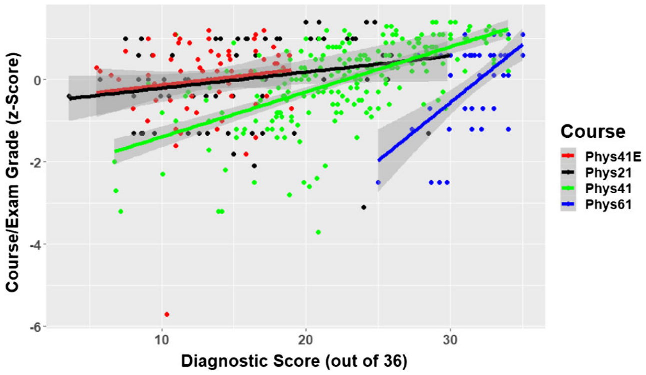

FIG. 2. Student course grades (Phys 61 and Phys 21) or exam grades (Phys 41 and Phys 41E) as a function of diagnostic exam scores. Points in red are for Phys 41E (the calculus-based course for students with no prior physics), points in black are for Phys 21 (the algebra-based mechanics course), points in green are for Phys 41 (the calculus based course for scientists and engineers), and points in blue are for Phys 61 (the honors course for physics majors). We include ordinary least squares trend lines and associated standard errors to guide the reader's eye.

TABLE II. Regression of Phys 61 course grades on measures of incoming preparation. $^{***}p < 0.001$.

Phys 61 course grade	Model 1	Model 2
SAT score		0.056 (0.14)
Diagnostic score	0.58^{***} (0.11)	0.57^{***} (0.12)
FMCE prescore		Not applicable
R squared	0.34	0.34

TABLE V. Regression of Phys 21 course grade on measures of incoming preparation. $^{\dagger}p < 0.10$, $^{***}p < 0.001$.

Phys 21 course grade	Model 1	Model 2
SAT score		0.43^{***} (0.095)
Diagnostic score	$0.20^{\dagger}$ (0.11)	0.13 (0.11)
FMCE prescore		Not applicable
R squared	0.052	0.24

with each standard deviation increase in diagnostic score corresponding to a 0.58 standard deviation increase in course grade. The total variance explained by diagnostic score is 34% (model 1). We find that SAT scores explain no additional variance in scores after controlling for diagnostic scores. We note that there is a ceiling effect in SAT scores for this population—the average SAT percentile score is 100% with a standard deviation of 0.73%.

The regression results for Phys 41 are in Table III. We find that the diagnostic score is a strong predictor of both course grade (model 1a) and exam grade (model 1b), but that the diagnostic predicts more than twice as much of the variation in exam scores (39%) as it does for course grades (18%). This is not surprising as the translation between scores on exams and homework into course grades is very nonlinear in a normal year, with a heavy weighting of A’s, and because this was the term interrupted by COVID, the grading policies were made more liberal. We find that both SAT score and diagnostic score are significant predictors of course grade (model 2a), but that adding in SAT scores does not explain any additional variance in course grades. This indicates that SAT scores and diagnostic scores are strongly correlated. We find that all three measures of incoming preparation are significant predictors of exam grades (model 2b), but that adding in FMCE scores and SAT scores only adds an additional 7% of the variation explained in exam grades. We note that the *R* squared in model 1b is slightly higher than the total *R* squared reported in Ref. [9], indicating that the diagnostic is a better predictor of exam performance than SAT scores and FMCE scores combined.

The regression results for Phys 41E are reported in Table IV. We find that the diagnostic is a poor predictor of performance in Phys 41E, explaining 3.5% of the variation in course grades (model 1a) and 1.3% of the variation in exam grades (model 1b). When we add in the other measures of incoming preparation, we find that SAT score is the best predictor of course performance, but still relatively weak (*R* squared less than 10%). We note that Phys 41E was designed so that student outcomes should be independent of students’ incoming physics preparation.

The regression results for Phys 21 are reported in Table V. We find that diagnostic score is a marginal predictor of course grade (explaining 5.2% of the variation), and that SAT score is a much stronger predictor of performance in Phys 21, explaining an additional 20% of the variation in course grades.

TABLE III. Regression of Phys 41 course and exam grades on measures of incoming preparation. $^{**}p < 0.01$, $^{***}p < 0.001$.

Phys 41 outcomes	Model 1a (grade)	Model 1b (exam)	Model 2a (grade)	Model 2b (exam)
SAT score			0.16^{**} (0.060)	0.25^{***} (0.060)
Diagnostic score	0.37^{***} (0.052)	0.61^{***} (0.05)	0.27^{**} (0.094)	0.40^{***} (0.068)
FMCE prescore			0.083 (0.080)	0.18^{**} (0.059)
R squared	0.18	0.39	0.18	0.46

TABLE IV. Regression of Phys 41E course and exam grades on measures of incoming preparation. $^{*}p < 0.05$.

Phys 41E outcomes	Model 1a (grade)	Model 1b (exam)	Model 2a (grade)	Model 2b (exam)
SAT score			0.22 (0.13)	0.29^{*} (0.12)
Diagnostic score	0.18 (0.14)	0.11 (0.12)	0.14 (0.14)	0.058 (0.12)
FMCE prescore			−0.0074 (0.13)	0.047 (0.13)
R squared	0.035	0.013	0.095	0.10

TABLE VI. Difficulty (fraction of students getting a question correct) and discrimination (correlation with course grade) for each item on the diagnostic. Bolded questions indicate good questions for sorting students into different courses, underlined questions are good for distinguishing honors students. Italicized questions are predictive of performance in at least one of the courses.

Question	Discrimination (Physics 21)	Difficulty (Physics 21)	Discrimination (Physics 41E)	Difficulty (Physics 41E)	Discrimination (Physics 41)	Difficulty (Physics 41)	Discrimination (Physics 61)	Difficulty (Physics 61)
1. Vector Subtraction	***0.22***	***0.30***	***0.02***	***0.14***	***0.11***	***0.56***	***0.33***	***0.90***
2. Unit Vector	***0.063***	***0.33***	***−0.01***	***0.22***	***0.29***	***0.52***	***0.53***	***0.90***
3. Vector Magnitude	***0.35***	***0.46***	***0.15***	***0.49***	***0.17***	***0.79***		***1.00***
4a. $A \cdot B = 0?$	*0.22*	*0.66*	*0.17*	*0.62*	*0.24*	*0.76*	*0.37*	*0.96*
4b. $A \times B = 0?$	*−0.016*	*0.67*	*−0.14*	*0.67*	*0.18*	*0.79*	*0.35*	*0.98*
4c. $A \times A = 0?$	0.077	0.30	0.17	0.28	0.19	0.42	0.048	0.92
4d. $A \cdot (A \times B) = 0?$	*0.23*	*0.45*	*0.19*	*0.51*	*0.03*	*0.59*	*0.18*	*0.87*
5. Net force on a box	0.1	0.81	0.02	0.61	0.08	0.91	−0.086	0.98
6. System of two equations	0.052	0.70	0.14	0.80	0.10	0.87	−0.046	0.94
7. Taylor expansion analytical	*−0.13*	*0.16*	*−0.09*	*0.29*	*0.11*	*0.29*	*0.34*	*0.65*
8a. Taylor expansion graph	*0.19*	*0.54*	*0.18*	*0.42*	*0.22*	*0.67*	*0.35*	*0.98*
8b. Taylor expansion graph	*0.057*	*0.49*	*0.00*	*0.46*	*0.20*	*0.58*	*0.21*	*0.90*
8c. Taylor expansion graph	*−0.014*	*0.34*	*−0.06*	*0.30*	*0.17*	*0.40*	*0.38*	*0.79*
9. Derivative chain rule	*0.21*	*0.70*	*0.24*	*0.64*	*0.11*	*0.81*		*1.00*
10. Basic integral	*0.046*	*0.69*	*0.22*	*0.74*	*0.04*	*0.86*	*0.14*	*0.92*
11. Wavelength of function	*−0.011*	*0.31*	*0.13*	*0.26*	*0.29*	*0.36*	*0.27*	*0.81*
12. Dimensional analysis	**−0.029**	**0.13**	**0.10**	**0.17**	**0.07**	**0.38**	**0.16**	**0.79**
13. $X = v\Delta tX = ½at^2X = ½at^2$	−0.2	0.70	−0.14	0.67	0.04	0.92	−0.0089	0.98
14. Relative velocity	0.015	0.28	0.11	0.28	0.12	0.36	0.17	0.75
15. Acceleration from $\boldsymbol{x(t)}$ ***graph***	***−0.13***	***0.52***	***0.13***	***0.28***	***0.22***	***0.72***	***0.23***	***0.92***
16. Balance beam static	***0.21***	***0.52***	***−0.13***	***0.55***	***0.17***	***0.81***		***1.00***
17. Normal component of gravity	*0.34*	*0.49*	*0.03*	*0.43*	*0.12*	*0.64*	*0.087*	*0.90*
18. Complementary angles	***−0.071***	***0.46***	***−0.33***	***0.35***	***0.29***	***0.76***	***−0.086***	***0.98***
19. Tangential component of gravity	**−0.069**	**0.42**	**0.16**	**0.30**	**0.14**	**0.58**	**0.048**	**0.92**
20. $\boldsymbol{X = ½\,at^2}$	***0.082***	***0.30***	***0.07***	***0.13***	***0.25***	***0.55***	***0.094***	***0.98***
21. Energy	**−0.0074**	**0.52**	**−0.09**	**0.33**	**0.18**	**0.72**	**0.19**	**0.96**
22. Energy	−0.079	0.79	−0.08	0.64	0.07	0.97	0.094	0.98
23. Projectile Motion (two arcs same height)	**−0.02**	**0.40**	**−0.10**	**0.32**	**0.02**	**0.63**	**0.19**	**0.96**
24. Acceleration from v(t) graph	0.11	0.67	−0.11	0.59	0.20	0.93		1.00
25. Velocity—position relationship	*0.23*	*0.60*	*−0.02*	*0.49*	*0.10*	*0.85*	*−0.091*	*0.94*
26. Time to fall independent of mass	−0.016	0.64	−0.13	0.61	0.13	0.81	0.19	0.96
27. $\boldsymbol{X = ½\,at^2}$	***0.048***	***0.22***	***0.12***	***0.10***	***0.27***	***0.55***	***0.014***	***0.94***
28. Newton's 3rd law collision	***0.28***	***0.45***	***−0.04***	***0.22***	***0.15***	***0.68***	***0.087***	***0.90***
29. Newton's 3rd law collision	**0.1**	**0.36**	**0.10**	**0.20**	**0.19**	**0.51**	**0.025**	**0.73**

(Table continued)

TABLE VI. *(Continued)*

Question	Discrimination (Physics 21)	Difficulty (Physics 21)	Discrimination (Physics 41E)	Difficulty (Physics 41E)	Discrimination (Physics 41)	Difficulty (Physics 41)	Discrimination (Physics 61)	Difficulty (Physics 61)
30. Newton's 3rd law collision	***0.31***	***0.39***	***0.11***	***0.14***	***0.12***	***0.63***		***1.00***
31. ID Force 1	*−0.13*	*0.75*	*0.33*	*0.75*	*0.02*	*0.85*	*−0.086*	*0.98*
32. ID Force 2	−0.042	0.87	−0.11	0.86	0.10	0.97		1.00
33. ID Force 3	0.1	0.78	−0.06	0.83	0.04	0.89	0.17	0.98
34. ID Force 4	*−0.049*	*0.49*	*0.12*	*0.49*	*0.26*	*0.74*	*0.31*	*0.88*
35. Uniform circular motion	*−0.15*	*0.25*	*−0.03*	*0.20*	*0.17*	*0.36*	*0.27*	*0.79*
36. Angular momentum	−0.18	0.04	−0.14	0.12	0.05	0.18	0.04	0.63
37. Electrostatic force	***0.37***	***0.33***	***0.15***	***0.32***	***0.24***	***0.68***	***0.13***	***0.92***
38. Magnetic field wire	***0.14***	***0.25***	***0.14***	***0.17***	***0.22***	***0.42***	***−0.14***	***0.87***

IV. ITEM-LEVEL ANALYSIS

In this analysis, the difficulty of a question was defined as the fraction of students in each course who got it correct, and the discrimination of a question was defined as the correlation between students' responses to the question and their course grades (see Table VIII). We judged an individual question to be good at sorting students into the appropriate courses if there was a substantial difference between the difficulty for Phys 61, Phys 41, and Phys 41E or 21. We set a threshold difference of 0.2, meaning that if Phys 61 students got a question correct 20% more often than Phys 41 students, who in turn got the question correct 20% more than 41E or 21 students, the question was judged to be good at sorting between the courses. We also identified questions which were good at identifying students for Phys 61 only. To do this, we set a threshold of Phys 61 students getting the question right at least 30% more than all other students, and there being less than a 20% difference between all the other courses.

We note that we have used which course a student enrolled in instead of their relative performance on the diagnostic to divide students into categories because it provides useful information on what is and is not measured in particular courses. This is particularly true for the discrimination coefficients. Furthermore, the scatterplot in Fig. 2 reveals that dividing students into courses is a very good proxy for their relative performance on the diagnostic, thus we expect the confounding effects of which course a student chose to be small.

Using this procedure, we identified 17 questions that were good for sorting students into different courses and 9 questions that were good for distinguishing Phys 61 students from the rest of the population. The questions that were good for distinguishing the honors students covered Taylor expansions, wavelengths of sinusoidal functions, relative velocity in 2D, and rotational motion. The questions that were good for sorting all students into different courses covered basics of vectors (addition, unit vectors, magnitude), dimensional analysis, static equilibrium, calculating components of forces, projectile motion, the relationship between displacement and acceleration, Newton's 3rd law for collisions, and two questions that covered basic ideas from electricity and magnetism.

We label any question with a discrimination coefficient of 0.2 or greater as a "good predictor" of performance in a given course. We found 11 questions which were predictive of performance in Phys 21, 3 questions which were predictive of performance in Phys 41E, 11 questions which were predictive of performance in Phys 41, and 12 questions which were predictive of performance in Phys 61. 15 questions were not predictive of performance in any course.

The best predictors of performance in Phys 21 were questions that asked students to identify the normal component of the force of gravity on an object sitting on an inclined plane, that asked students to compute the magnitude of a vector, and that asked students to identify the magnitude of an electrostatic force when the distance between two point charges was doubled. The questions predictive of performance in Phys 41E were basic calculus questions (derivative chain rule and basic integration), and a question that asked students to identify complementary angles. The questions most predictive of performance in Phys 41 were the same complementary angles question, a question asking students to identify the wavelength of a sinusoidal function, and a question asking students to identify a unit vector for a given arbitrary vector. The questions most predictive of performance in Phys 61 were the unit vector question, a question about the vector dot product, and a question about Taylor expansion.

There were 8 questions on the diagnostic which were not good predictors of performance in any of the courses and were also not good for sorting students into different courses based on the criteria we describe above: Questions 5, 6, 13, 22, 24, 26, 32, and 33. These questions cover a range of topics, but they are all quite easy for this population—at least 60% of the least well-prepared students got these questions correct. Despite not being identified as "good" questions by our algorithm, it is clear from Table VI that these questions may still be valuable to include. For example, question 22 is very good at distinguishing students should be in Phys 41E from others (students who get this wrong should enroll in Phys 41E). Similarly, question 26 shows reasonable discrimination between the different courses (difference in difficulty > 0.15) but did not meet the threshold we set.

We did a factor analysis of the exam results. The scree plot showed that all questions were heavily loaded onto a single factor that explained more of the variance, with all other factors contributing very little. This suggests that we met our design goal of avoiding any redundancy between questions in order to obtain the maximum discrimination with the minimum number of questions.

VI. DISCUSSION

The diagnostic is a strong predictor of performance in both Phys 61 and 41—explaining 35% of the variation in Phys 61 course grades and 40% of the variation in Phys 41 exam grades. Notably, the diagnostic is a much better predictor of exam grades than course grades, for the reasons discussed above. It is a modestly better predictor by itself than the combination of SAT scores and FMCE scores [9]. The diagnostic is a poor predictor of performance in Phys 21, while SAT scores are an important predictor. We expect that this is due to the additional content covered in Phys 21 beyond what is covered on the diagnostic. Over half of the ten-week course is spent on the topics of heat, fluids, and waves, which are not probed by the diagnostic. We expect that in an algebra-based course covering only mechanics, the diagnostic would be a good predictor. The diagnostic is also a poor predictor of performance in Phys 41E, as are all measures of incoming preparation. This is likely due to the design of Phys 41E. It is intended to be a course where performance is independent of prior preparation. This analysis suggests that the instructors of that course were successful in achieving their course design goals.

We note that, though the amount of variation in exam scores or course grades explained by the diagnostic is high by social science standards, it is far from 100%. This implies that instructors should be careful in using strict cutoff scores for placing students. Though most students in our samples adhered to the advice provided to them, there were certainly many students who performed below a threshold score that performed well in a given course and vice versa. Indeed, the diagnostic exam fails to predict nearly $2/3$ of the variation in final exam scores, suggesting there are many other factors other than prior preparation which may impact student success in these courses. However, the diagnostic will still be able to help students choose a course that is challenging, but in which they will have a high probability of success.

We found 17 individual questions that were particularly good at sorting students into various courses, which covered a range of topics including kinematics, dimensional analysis, vectors, Newton's 3rd law, forces and torques, and topics from electricity and magnetism. We found an additional 9 questions that were particularly discriminating with regard to student performance in honors physics. Those questions covered Taylor expansions, rotational motion, and other more advanced mathematical ideas.

We found 27 questions to be strongly predictive of performance in at least one of the courses we studied here. Some of these correlations reflect content overlap with a particular course, such as the Newton's 3rd law questions about collisions in Phys 21. However, there are some questions that we expect are proxies for students' general physics preparation. In particular, the questions about unit vectors and function wavelengths are good at predicting performance in Phys 61 and 41, but these ideas are not directly used in those courses. These questions are likely measuring the deeper understanding of mathematical functions and vectors, which translates into better course performance.

The select population studied here is the most notable limitation to the generalization of our analysis—indeed the distribution of student math SAT scores is quite narrow and has a high mean. However, the distribution of scores on the FMCE precourse test has a substantial fraction close to zero [19]. We do see that this diagnostic is able to discern a wide range of prior physics knowledge among this population. Furthermore, many of the students in this population have never taken physics, so we would not expect their diagnostic performance to be that different from many high school students who have never taken physics. It would be good to test those questions on the diagnostic that nearly all students answer correctly to see if the results are similar in other populations.

VII. CONCLUSIONS

In this work we described the validation of a diagnostic test for physics placement that was developed at Stanford University. The diagnostic test covers a broad range of topics in both physics and mathematics relevant to introductory mechanics. The diagnostics is a good predictor of student success in standard mechanics courses and is able to appropriately recommend which level of physics course a student should consider taking when they enter the university. It also identifies a number of areas that are particularly important to student success, and hence can provide useful guidance and assessment on improving instruction. We hope that physics departments at other universities may use this diagnostic where relevant and will conduct similar analyses to determine its validity and value.

[1] E. Hsu and D. Bressoud, Placement, and student performance in calculus I, in *Insights and Recommendations from the MAA National Study of College Calculus*, edited by D. Bressoud, V. Mesa, and C. Rasmussen (MAA Press, Washington, D.C., 2015), Chap. 5.

[2] C. McFate and J. Olmsted III, Assessing student preparation through placement tests, J. Chem. Educ. **76** (1999).

[3] G. W. White, D. M. Miller, L. C. Matten, D. C. Englert, and M. Douglas Scott, National and local proficiency tests: Their validity for an introductory biology course, Educ. Psych Meas. **36**, 993 (1976).

[4] R. L. Forrest, D. W. Stokes, A. B. Burridge, and C. D. Voight, Math remediation intervention for student success in the algebra-based introductory physics course, Phys. Rev. Phys. Educ. Res. **13**, 020137 (2017).

[5] A. G. Medhanie, D. N. Dupuis, B. LeBeau, M. R. Harwell, and T. R. Post, The role of the ACCUPLACER mathematics placement test on a student's first college mathematics course, Educ. Psychol. Meas. **72**, 332 (2012).

[6] J. Scott-Clayton, Do High-Stakes Placement Exams Predict College Success?, CCRC Working Paper No. 41 (2012).

[7] M. J. Legg, J. C. Legg, and T. J. Greenbowe, Analysis of success in general chemistry based on diagnostic testing using logistic regression, J. Chem. Educ. **78**, 1117 (2001).

[8] W. K. Adams and C. E. Wieman, Development and validation of instruments to measure learning of expert-like thinking, Int. J. Sci. Educ. **33**, 1289 (2011).

[9] S. Salehi, E. W. Burkholder, G. P. Lepage, S. Pollock, and C. Wieman, Demographic gaps or preparation gaps?: The large impact of incoming preparation on performance of students in introductory physics, Phys. Rev. Phys. Educ. Res. **15**, 020114 (2019).

[10] The diagnostic may be accessed here: https://stanforduniversity.qualtrics.com/jfe/form/SV_6sdi9NK9c7ynWx7.

[11] The College Board(2018), Guide to the 2018 ACT/SAT concordance, Retrieved from: https://collegereadiness.collegeboard.org/pdf/guide-2018-act-sat-concordance.pdf.

[12] R. Thornton and D. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula, Am. J. Phys. **66**, 338 (1998).

[13] D. Kleppner and R. Kolenkow, *An Introduction to Mechanics*, 2nd ed. (Cambridge University Press, Cambridge, England, 2013).

[14] D. Morin, *Special Relativity: For the Enthusiastic Beginner*, 1st ed. (Create Space, Cambridge, MA, 2017).

[15] H. D. Young and R. A. Freedman, *University Physics with Modern Physics*, 14th ed. (Pearson, Boston, 2015).

[16] L. C. McDermott and P. S. Shaffer, *Tutorials in Introductory Physics* (Prentice Hall, Upper Saddle River, NJ, 2002).

[17] D. C. Giancoli, *Physics: Principles with Applications*, 7th ed. (Pearson Boston, 2014).

[18] J. Nissen, R. Donatello, and B. Van Dusen, Missing data and bias in physics education research: A case for using multiple imputation, Phys. Rev. Phys. Educ. Res. **15**, 020106 (2019).

[19] E. W. Burkholder and C. E. Wieman, What do AP physics courses teach and the AP physics exam measure?, Phys. Rev. Phys. Educ. Res. **15**, 020117 (2019).