# Pre-instruction diagnostic tests can predict grade probabilities in introductory physics

David E. Meltzer [a)]
School of Applied Sciences and Arts
College of Integrative Sciences and Arts
Arizona State University
Mesa, AZ 85212

Dakota H. King
University of Arizona College of Veterinary Medicine
Oro Valley, Arizona 85737

**Editor's Note**: This large-scale study (2141 students across 31 classes at 5 institutions with 8 instructors) shows that students who score in the top and bottom quartiles of pretests are much more likely to earn grades in the top and bottom quartiles, respectively, compared to the chances that they will move between these quartiles. Physics instructors will want to consider the implications of these results.

**Abstract**

We have investigated the probabilities of earning high (top-quartile) and low (bottom-quartile) course grades in introductory university physics courses for students in two different groups: one, those who scored in the top quartile of their class on one of three diagnostic pretests, and the other composed of those who scored in the bottom quartile on the same test. The tests employed were the Force Concept Inventory (a physics concept test), the Lawson Test of Scientific Reasoning, and a newly developed mathematics test that includes only pre-college mathematics questions; all pretests were administered before or near the beginning of the course. Our investigation includes over 2000 students enrolled in 31 introductory physics classes taught by eight instructors at four universities. We found with 97% consistency that top-quartile scorers on any of the pretests were more likely to get high (top-quartile) grades and less likely to get low (bottom-quartile) grades than were bottom-quartile scorers on the same pretest. Top-quartile scorers on the pretests were, on average, four to six times as likely to receive high grades, and one-third to one-fifth as likely to receive low grades, compared to bottom-quartile scorers on the same pretests. These results are consistent with mathematical models of empirical data published by Salehi et al. (Phys. Rev. Phys. Educ. Res. **15**(2), 020114 (2019)) and with their cautions regarding the potentially serious implications of these findings for the careers of poorly prepared college physics students.

## I. INTRODUCTION

For more than 90 years, researchers have investigated factors associated with student success in introductory college physics courses. Before 1955, such studies were rare.[1,2] From the late 1950s through the 1980s there were many investigations, most of which focused on mathematical and reasoning skills; students' physics knowledge was sometimes examined as well.[3,4,5,6,7,8,9,10,11,12,13,14,15,16] Typically, researchers would determine the correlation between college physics students' final course or exam grades and their scores on various tests given before or at the beginning of instruction. Standardized tests such as the Scholastic Aptitude Test (SAT) or American College Testing (ACT) were sometimes used, while some investigators devised their own mathematics diagnostic tests.[17,18,19] The reasoning skills tests were generally not standardized and often were assembled in ad hoc fashion by researchers drawing on test items (or variations) that had been developed by other investigators. With the 1992 publication of the mechanics concept test known as "Force Concept Inventory" (FCI), researchers gained another tool that was sometimes employed to assess students' physics concept knowledge.[20]

In more recent years, the number of potentially influential factors examined has dramatically increased. A variety of demographic variables have been explored, including for example race, ethnicity, gender, and family educational background.[21,22,23] Other potential factors occasionally examined include self-efficacy and other affective factors, spatial reasoning ability, "sense of belonging," and students' previous experience (or lack of it) in high school physics and math courses.[24,25,26,27,28,29,30] The explanation for the continued interest in this topic is straightforward: If physics instructors can better understand the factors that lead to success in physics—and the obstacles that get in the way—they might be better able to guide and prepare students in a manner leading to improved performance.

A separate route of investigation has focused on the relation between pretest scores and measures of physics *learning* rather than mere performance on grades or final exams. These studies typically assessed students' physics knowledge both before and at the end of instruction to arrive at some measure of what the students had learned in the process.[31,32] This approach lies outside of our current focus and we won't address it further in this paper. However, several of these studies have special relevance to our work and require discussion. A key role was played by Coletta and Phillips and their collaborators who, in a series of papers beginning in 2005, reported that students' pretest scores on the Lawson Test of Scientific Reasoning (LTSR) had a significant correlation with their physics learning gains.[33,34,35,36] (The LTSR was developed by A.E. Lawson in 1978; we will discuss it further in Section II below.[37]) Dubson and Pollock reported confirmation of these findings in 2006 and added an important one of their own, that is: pretest scores on the LTSR also had a strong correlation with students' final course *grades* in an introductory mechanics course.[38] In a follow-up published in 2008, Pollock reported the same LTSR-grade correlation but this time for the second semester of the introductory course that focused on electricity and magnetism.[39] We have incorporated the raw data acquired by Dubson and Pollock into our study and so include both of their courses in our own data sample and analysis.

It would require many pages and take us too far afield to discuss in any detail the results of the dozens of other investigations alluded to above. We can succinctly summarize the findings of those which are most directly relevant to the present study by saying that nearly all of them found positive and generally significant correlation between students' physics course grades and

scores on pre-instruction tests of mathematics skill, reasoning ability, and physics concept knowledge. The magnitude of the correlation coefficients varied widely, generally in the range 0.20-0.50, depending on course, instructor, institution, and many other variables; few consistent patterns were observed. In almost no cases were concrete predictions of actual course grades examined except in the context of goodness-of-fit measures applied to a statistical model or correlation coefficients associated with specific grades. An exception is the work of Halloun and Hestenes (Ref. 14). These authors devised a predictive model that incorporated scores on both math skills and physics concept knowledge pretests and used it to predict students' final letter grades. They reported that about 40% of the students fell into a "low competence" category based on their scores on the predictive model, and about 95% of students in this category received course grades of C or less.

The problem with using letter grades as an analytical criterion is that methods of assigning such grades vary very widely among institutions, instructors, and courses. A more generalized criterion is to examine particular segments of the student population and compare their course outcomes with those of other segments of the same population. It is natural to focus on the students at greatest potential risk—that is, those who rank in the lower part of the class—and the most revealing comparison would be with those ranked toward the top of the class. This "bottom vs. top" comparison is most likely to reveal any existing patterns between pre-instruction measures and final course outcomes. The existence and scale of differences in the course outcomes might then provide some guidance as to the severity of the problem and the desirability of acting on the pre-instruction measures. One might, for example, divide the class into a top, middle, and bottom third, or—for a more dramatic comparison—into quartiles, comparing top and bottom quartile. That latter choice was made in a recent research study by Salehi et al. (Ref. 21) and we have adopted that same approach here. (It is worth noting that bottom-to-top comparisons may reveal clear patterns even when a linear correlation analysis yields null or ambiguous results, as we discuss in more detail in Section III.)

Salehi et al. developed predictive models that incorporated pre-instruction scores on math skills and physics concept tests, determining measures of correlation (adjusted $R$-squared or $R^2_{\text{adj}}$) between predicted and actual grades on final exams in an introductory mechanics course. They created a mathematical model by making certain assumptions about the nature of the data and then calculated that the values of $R$-squared that they had found implied that bottom-quartile scorers on the predictive model were four times more likely to obtain bottom-quartile grades than peers who scored in the top quartile on the predictive model.

The investigation we report here constitutes, in part, a test of Salehi et al.'s prediction. We examined scores obtained on three pre-instruction diagnostic tests by students enrolled in introductory physics courses; the three tests assessed reasoning ability, mechanics concept knowledge, and skills in pre-college mathematics. We focused solely on those students whose scores placed them in either the top quartile (ranking in the top 25% of their class) or bottom quartile (bottom 25% of their class) on the various pretest measures. We then determined empirically the probability that students in each of these two groups received a final course grade that ranked them either in the top quartile (top-25% grade) or bottom quartile (bottom 25%) of their class. Finally, we compared the high- and low-grade probabilities of the two pretest scorer groups (top and bottom quartile) to each other. We found dramatic differences in those probabilities that implied that high scorers on the pretests were much more likely to get high

grades, and much less likely to get low grades, than low scorers—by factors of two to six. This appears to be the first concrete report of such grade comparisons since the 1985 work of Halloun and Hestenes, and is in broad agreement with the predictions made by Salehi et al. We describe the details of the investigation below and in the final section discuss the instructional implications of the findings by Salehi et al. in light of our own empirical observations.

## II. SAMPLE AND METHOD

### A. Sample

Our population sample consisted of 31 distinct classes taught by eight different instructors at four universities; over 2000 total students were enrolled in these courses. The earliest course for which we have data took place in 2005; the most recent in 2024. All courses were part of the standard introductory university physics sequence, including both algebra- and calculus-based courses and both the first and second semester of the standard sequence. Nearly all of the courses incorporated research-based, active-learning techniques to a significant degree; however, each of the instructors had their own very distinctive approaches to implementing these methods.[40] The universities included Arizona State University, both the Polytechnic (ASU-P) and Tempe (ASU-T) campuses, Loyola Marymount University (LMU), the University of Colorado at Boulder (CU), and the University of West Florida (UWF). (See Table 1.) With the cooperation of the course instructors, we obtained detailed grade and pretest-score data, including the exact number of final grade points (not just letter grades) obtained by each student. The student sample for each class was restricted to those students for whom both final grade points and pretest scores were available, and we focused our investigation on students whose final grade points ranked in either the top or bottom quartile of their class.

| |
|---|
| **Course codes** |
| Alg-1: Algebra-based course, first semester |
| Alg-2: Algebra-based course, second semester |
| Calc-1: Calculus-based course, first semester |
| Calc-2: Calculus-based course, second semester |
| **Institution codes** |
| ASU-P: Arizona State University, Polytechnic campus |
| ASU-T: Arizona State University, Tempe campus |
| LMU: Loyola Marymount University |
| UWF: University of West Florida |
| CU: University of Colorado, Boulder |
| **Pretest codes** |
| Math: Mathematics Diagnostic Test (administered online in all cases) |
| Lawson: Lawson Test of Scientific Reasoning (administered online at ASU-P and in Alg-1 2005 CU, on paper in Calc-2 2007 CU and at LMU) |
| FCI: Force Concept Inventory (administered online at UWF and on paper at ASU-P, LMU, and CU) |

**Table I**. Course, institution, and diagnostic pretest abbreviations used in this paper.

### B. Diagnostic measures

We used three multiple-choice diagnostic pretests: (1) the Force Concept Inventory (FCI), a 30-item test of mechanics conceptual knowledge; (2) the Lawson Test of Scientific Reasoning (LTSR, hereafter referred to as "Lawson"), a 24-item diagnostic that includes questions on correlational, proportional, probabilistic, and control-of-variables reasoning; (3) a 16-item mathematics diagnostic test ("Math") that includes questions on trigonometry, algebra, graphing, and geometry; all questions on the Math test were at the pre-college level and calculators were allowed. Students were either required to take the pretests or offered small amounts of participation points if they took them voluntarily. Both the Lawson and FCI tests are available to instructors at PhysPort.org;[41] the Math test is provided in the Supplementary material.

The mathematics test that we employ is new and was developed through our separate investigation into mathematical difficulties encountered by introductory physics students. Available online and including only 16 multiple-choice items, most of which allow rapid responses, it is easy to administer and requires little of students' time (typically 15-20 minutes), giving it advantages over instruments used in previous studies. During our development of this test, we compared results with those obtained from the in-class paper version that we had administered to thousands of students in the same classes at the same institution. We found that scores on nearly all items were very nearly the same between the two versions, but *only* if we excluded scores on online tests submitted less than five minutes after the students had started. Although we have not made the same kind of comparison between the online and paper versions of the Lawson test, we considered it prudent to apply the same constraint to those tests as well. Thus, our samples do not include Math or (for the most part) Lawson scores for tests that were submitted online after less than five minutes from the time the students began their work. (We lacked the timing data for the Lawson test administered at CU to implement the restriction in that case.)

### C. Determination of top- and bottom-quartile groups

#### *1. Grades*

We define Q as 25% of the students. If the number of students was not divisible by 4, then Q is not an integer. In that case, the quartile cutoffs were adjusted to include the smaller (integer) number of students. For example, in a class of 42 students, Q = 10.5, so the upper quartile included students 1-10 and the lower quartile included students 33-42. Since we had exact grade points (to three or more significant figures) for each student, there was never any ambiguity of precisely where to set the high- and low-grade cutoffs.

#### *2. Pretest scores*

The determination of top- and bottom-quartile groups among the pretest scorers was complicated by the fact the multiple students in most samples shared the same scores on the pretests. This was inevitable, given the integer scoring with maximum scores of 16, 24, and 30 for the Math, Lawson, and FCI tests respectively. Thus the procedure we adopted to define the top-quartile scorers was first to specify groups as follows: (1) Define $Q$ as 25% of the number of students in the full sample; (2) Define a first group that includes all students whose scores put them definitely in the top quartile, and use $F \leq Q$ to represent the number of such students; (3) in cases where $F<Q$, define a second group as those students whose pretest scores were one point

lower than the lowest of those in the first group, and define $S$ as the number of students in this second group; (4) define $f=(Q-F)/S$. This fraction $F$ will be used in the calculations in Section D below.

**D. Determination of grade probabilities and odds ratios**

When determining the percentage of top-quartile test scorers who earn top-quartile grades, the second group of students is weighted by the fraction $f$. If $T$ students out of $F$ earn grades in the top quartile, while $t$ students out of $S$ earn grades in the top quartile, then we use $T + ft$ as the total number of students from the top quartile of pretest scores whose grades were also in the top quartile. As an example, let's again assume that there are 42 students in the full sample so that $Q$ = 10.5. Suppose now that the top eight scorers on the Math pretest have scores of 14 or higher, while five students obtained a score of 13. Then we have $Q$ = 10.5, $F$ = 8, $S$ = 5, and $f$ = (10.5−8)/5 = 0.5. Now suppose that four (out of eight) students in the first group obtained top-quartile grades ($T$=4) while two (out of five) students in the second group did so ($t$=2). Then we would say that the number of top-quartile grades obtained by students with top-quartile Math scores is equal to $T+ft$ = 4+(0.5)(2)=5. The probability $P$ of obtaining a top-quartile grade is then given by $P=(T+ft)/Q$ = (4+1)/10.5=48% in this example. (A figure that illustrates this example is provided in the Supplementary Material.) This procedure, although somewhat complicated, allows for consistent and objective determination of top- and bottom-quartile groups for both grades and pretest scores.

We followed an analogous procedure to determine the probability of bottom-quartile grades from the top-quartile scorers, as well as those of top- and bottom-quartile grades from the bottom-quartile scorers. Next, we determined the probability ratios (the "odds ratios") as the ratios of these probabilities. For example, in Table II, the "High grade odds ratio" is given by [(probability of a top-quartile Lawson scorer obtaining a top-quartile grade) ÷ (probability of a bottom-quartile Lawson scorer obtaining a top-quartile grade)]. In 25% of all cases examined, the odds ratio was undefined because the probability in the denominator of the ratio was 0%.

The average values on the last lines of the tables are unweighted averages of the probabilities of the individual classes contained in each pretest sample, and the average odds ratio is calculate from the ratio of these average probabilities. It is *not* the average of the odds ratios of the individual courses (many of which are undefined).

The unweighted average is the appropriate one to use in this case because the individual unit of study is the *class*, not the individual student. That is, we examine each class to determine the grade odds ratios for that class and we fully characterize the class by those ratios. The questions at issue, then, are (1) what is the magnitude of that ratio for a "typical" class, and (2) to what degree are those ratios are *consistent* across different classes of different size taught by different instructors at different institutions? We do not want to allow the size of the class to interfere with our assessment of the variance of the odds ratios across classes. For example, using a weighted average for a sample that includes a single class of $N$=469 when most classes have $N$ in the 30-50 range would vastly overrate the significance of the odds ratio of that one large class.

**E. Combined sample**

Most of the samples are small ($N$ < 50) and in most cases there are differences in instructor or class time and class size such that forming a combined sample (to increase sample size) is

methodologically highly questionable. However, there are three classes that were given at the same time of day at the same institution by the same instructor three years in a row, and that had roughly equal numbers of students enrolled. Thus we created a new sample by combining Alg-2 ASU-P 2022 with Alg-2 ASU-P 2023 and Alg-2 ASU-P 2024 to form the combined sample "Alg-2 ASU-P 2022-23-24." However, due to inevitable differences in population makeup and grading between the three courses, we did not simply merge the raw grade and pretest scores. Instead, we determined the percentile rank *within their own class* of each student on each of the measures (that is, grade points, Math score, and Lawson score) and merged those percentile values to form the combined sample. (We note that we also carried out a separate calculation using *z*-scores instead of percentile scores; it yielded similar results, and we do not consider it any further in this report.)

### F. Samples with multiple pretests

For classes in which two or more pretests were administered, we carried out a number of calculations, discussed below, in which we attempted to determine the relative degree of association of the different pretest scores with grade points. For all these calculations, the sample tested included *only* those students who had scores on both (or all three) of the pretests, as well as final course grades.

## III. RESULTS

### A. Grade probabilities

We found a consistently large difference in probability of receiving high (top-quartile) course grades between the high and low scorers (that is, top- and bottom-quartile scorers) on all three of the pretests, and a similar result for low (bottom-quartile) course grades. Calculated from the unweighted averages shown in the tables, we find that high scorers in a class were much more likely (by a factor of 4–6) to receive high grades, and much less likely (by a factor of 0.2–0.3) to receive low grades, than were low scorers in that class. This general pattern held for 113 out of 116 comparisons (97%) and for all four universities, although the quantitative range was large. In Tables II and III, we show the detailed results corresponding to the Lawson pretest; similar tables in the Appendix show the results for the Math and FCI pretests.

(We note that there is no apparent pattern to the three outliers among the 116 comparisons, as they were from three different instructors, all of whom had non-outlier values from the same or other pretests included in their sample.)

| Course | Campus | *N* | Top-quartile Lawson | Bottom-quartile Lawson | High-grade odds ratio |
|---|---|---|---|---|---|
| Alg-1 2021a | ASU-P | 37 | 49% | 11% | 4.3 |
| Alg-1 2021b | ASU-P | 36 | 41% | 11% | 3.7 |
| Alg-1 2022a | ASU-P | 41 | 49% | 10% | 5.0 |
| Alg-1 2022b | ASU-P | 53 | 58% | 10% | 5.8 |
| Alg-1 2023a | ASU-P | 36 | 39% | 33% | 1.2 |
| Alg-1 2023b | ASU-P | 43 | 55% | 10% | 5.5 |
| Alg-2 2022 | ASU-P | 66 | 52% | 4% | 11.9 |
| Alg-2 2023 | ASU-P | 76 | 51% | 16% | 3.2 |
| Alg-2 2024 | ASU-P | 90 | 41% | 5% | 8.0 |
| Alg-1 2005 | CU | 469 | 45% | 8% | 5.5 |
| Calc-2 2007 | CU | 276 | 57% | 8% | 6.9 |
| Alg-1 2007 | LMU | 24 | 50% | 0% | [undefined] |
| Alg-1 2009 | LMU | 51 | 34% | 11% | 3.2 |
| Alg-1 2011 | LMU | 57 | 53% | 18% | 2.9 |
| Alg-1 2012 | LMU | 44 | 64% | 6% | 10.5 |
| Alg-1 2013 | LMU | 30 | 53% | 12% | 4.6 |
| Alg-1 2014 | LMU | 33 | 61% | 0% | [undefined] |
| Alg-1 2015 | LMU | 24 | 63% | 0% | [undefined] |
| Alg-1 2016 | LMU | 35 | 41% | 0% | [undefined] |
| Alg-1 2018 | LMU | 47 | 54% | 9% | 6.3 |
| Alg-1 2021 | LMU | 27 | 44% | 0% | [undefined] |
| **AVERAGE (unweighted)** | | **[1595]** | **50%** | **9%** | **5.8** |

Table II. High-grade probability vs. Lawson pretest score. The columns show the percentages of students with top-quartile scores ("Top-quartile Lawson") and bottom-quartile scores ("Bottom-quartile Lawson") on the Lawson pretest who received top-quartile grades in their class. The High-grade odds ratio is found by dividing the value of the grade percentage of the top-quartile group by that of the bottom-quartile group. The bottom row shows total *N* [in brackets] and unweighted averages of the top- and bottom-quartile columns, while the ratio of those two averages is shown in the Odds Ratio column. (Course and campus code in Table 1.)

| Course | Campus | *N* | Top-quartile Lawson | Bottom-quartile Lawson | Low-grade odds ratio |
|---|---|---|---|---|---|
| Alg-1 2021a | ASU-P | 37 | 6% | 44% | 7.2 |
| Alg-1 2021b | ASU-P | 36 | 11% | 47% | 4.2 |
| Alg-1 2022a | ASU-P | 41 | 15% | 28% | 1.9 |
| Alg-1 2022b | ASU-P | 53 | 15% | 45% | 3.0 |
| Alg-1 2023a | ASU-P | 36 | 14% | 36% | 2.6 |
| Alg-1 2023b | ASU-P | 43 | 8% | 50% | 6.7 |
| Alg-2 2022 | ASU-P | 66 | 12% | 25% | 2.1 |
| Alg-2 2023 | ASU-P | 76 | 11% | 28% | 2.7 |
| Alg-2 2024 | ASU-P | 90 | 10% | 36% | 3.6 |
| Alg-1 2005 | CU | 469 | 10% | 42% | 4.4 |
| Calc-2 2007 | CU | 276 | 12% | 44% | 3.8 |
| Alg-1 2007 | LMU | 24 | 0% | 58% | [undefined] |
| Alg-1 2009 | LMU | 51 | 5% | 48% | 10.4 |
| Alg-1 2011 | LMU | 57 | 15% | 46% | 3.0 |
| Alg-1 2012 | LMU | 44 | 9% | 27% | 3.0 |
| ***Alg-1 2013*** | ***LMU*** | ***30*** | ***27%*** | ***12%*** | ***0.4*** |
| Alg-1 2014 | LMU | 33 | 0% | 68% | [undefined] |
| Alg-1 2015 | LMU | 24 | 0% | 75% | [undefined] |
| Alg-1 2016 | LMU | 35 | 11% | 46% | 4.0 |
| Alg-1 2018 | LMU | 47 | 16% | 42% | 2.7 |
| Alg-1 2021 | LMU | 27 | 0% | 89% | [undefined] |
| **AVERAGE (unweighted)** | | **[1595]** | **10%** | **45%** | **4.5** |

Table II. Low-grade probability vs. Lawson pretest score. The columns show the percentages of students with top-quartile scores ("Top-quartile Lawson") and bottom-quartile scores ("Bottom-quartile Lawson") on the Lawson pretest who received bottom-quartile grades in their class. The Low-grade odds ratio is found by dividing the value of the grade percentage of the bottom-quartile group by that of the top-quartile group. The bottom row shows total *N* [in brackets] and unweighted averages of the top- and bottom-quartile columns, while the ratio of those two averages (bottom divided by top) is shown in the Odds Ratio column. (Course and campus code in Table 1.) An outlier case (that does not follow the general pattern) is shown in bold red italics.

Another way to view the results is to compare students in specific pretest groups to a student selected at random. If selected at random, 25% of the students in a class would be expected to receive a course grade falling within any specified quartile. In contrast, we see that the actual proportions of the top- and bottom-quartile scorers receiving too- and bottom-quartile grades deviate from 25% by a wide margin in each case; see Table IV.

Table IV. Proportion of students receiving top- and bottom-quartile grades for a random group of students, for bottom-quartile scorers on each of the three diagnostic tests, and for top-quartile scorers on each of those tests.

| | **Top-quartile grade proportion** | **Bottom-quartile grade proportion** |
|---|---|---|
| **Random students** | 0.25 | 0.25 |
| **Bottom-quartile scorers on Lawson** | 0.09 | 0.45 |
| **Bottom-quartile scorers on Math** | 0.12 | 0.36 |
| **Bottom-quartile scorers on FCI** | 0.08 | 0.40 |
| | | |
| **Top-quartile scorers on Lawson** | 0.50 | 0.10 |
| **Top-quartile scorers on Math** | 0.44 | 0.12 |
| **Top-quartile scorers on FCI** | 0.47 | 0.11 |

When comparing top- and bottom-quartile scorers to each other, rather than to a random student, we find that top-quartile scorers on the pretests are four to six times as likely to receive high grades, and one-third to one-fifth as likely to receive low grades, compared to bottom-quartile scorers on the same pretests. The average probability ratio (high-scorer probability vs. low-scorer probability) for receiving a top-quartile grade was 5.8, 3.8, and 5.8 for the Lawson, Math, and FCI pretests, respectively. The average probability ratio (low-scorer probability vs. high-scorer probability) for receiving a bottom-quartile grade was 4.5, 3.1, and 3.7 for the Lawson, Math, and FCI pretests, respectively.

[Note that these figures are drawn from the unweighted averages shown in the tables. Had we instead weighted the averages according to class enrollment, some of the numbers would have been different.]

It is important to emphasize that these consistent quartile comparisons held even though the linear correlation coefficients between grades and pretest scores were not particularly high (most fell in the +0.3-0.6 range) and most scatterplots did not immediately reveal the underlying pattern. However, with closer examination, this pattern becomes evident in nearly all of the samples. As an illustration, we use the large, combined sample Alg-2 ASU-P 2022-23-24 and first show in Fig. 1 the grade odds ratios for this sample.

**Alg-2, ASU-P 2022-23-24 (*N*=216), Grade Odds Ratios**

| High-grade odds ratio [top-Q math/bottom-Q math] | High-grade odds ratio [top-Q Lawson/bottom-Q Lawson] |
|---|---|
| 2.8 | 4.1 |

| Low-grade odds ratio [bottom-Q math/top-Q math] | Low-grade odds ratio [bottom-Q Lawson/top-Q Lawson] |
|---|---|
| 4.8 | 2.4 |

Fig. 1. Grade odds ratios (high grade = top-quartile grade; low grade = bottom-quartile grade) for Math and Lawson pretests in the Alg-2 ASU-P 2022-23-24 sample. The high-grade odds ratio (top row) is the ratio of the proportion of high (top-quartile) scorers on the specified pretest (Math, left column; Lawson, right column) who had top-quartile course grades to the proportion of low (bottom-quartile) pretest scorers who also had top-quartile grades. For the Math pretest it was (40.7%/14.8%)=2.8, while for the Lawson pretest that ratio was (43.1%/10.5%) = 4.1. The low-grade odds ratio is the ratio of the proportion of low (bottom-quartile) pretest scorers who had bottom-quartile grades to the proportion of high (top-quartile) pretest scorers who had bottom-quartile grades. For the Math pretest it was (41.1%/8.6%)=4.8, while for the Lawson pretest, that ratio was (30.9%/13.0%) = 2.4.

In Fig. 2(a), the scatter plot of all grades and Math pretest scores ($r$ = +0.35) in this sample is shown; no pattern is readily discernable. However, in part (b) of that figure, only the points corresponding to top- and bottom-quartile grades are shown; here, a pattern is obvious: those with higher pretest scores had more top-quartile grades and fewer bottom-quartile grades than low scorers. Note that *all* Math scores of the top-and bottom grade quartiles are shown in this figure; the ovals are drawn to enclose roughly the top and bottom 50% of Math scores, respectively. The points in the top and bottom Math quartiles may be estimated by eye and seen to correspond approximately to the exact numbers in Fig. 1 which were calculated as described in II.D. (The varying separations between Math scores are artifacts introduced by the conversion to within-class percentiles and the differing numbers of students in each of the three courses combined in this sample.)

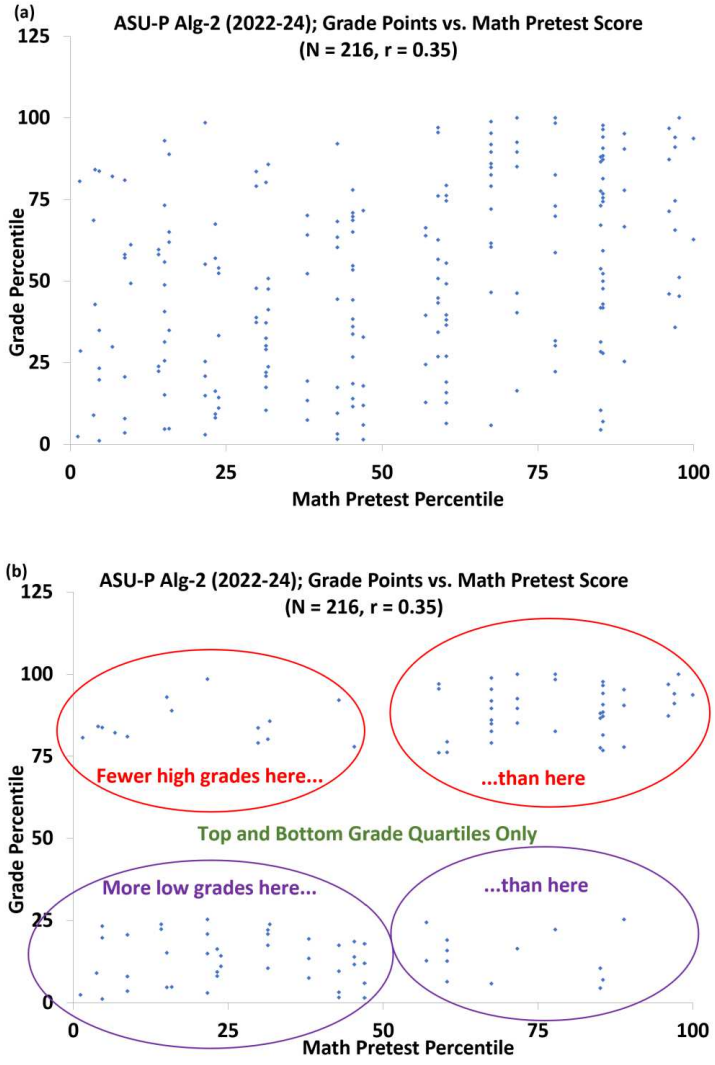



Fig. 2. (a) Scatterplot of grade percentile vs. Math pretest percentile for the combined sample ASU-P Alg-2 (2022, 2023, 2024). (b) Same data as (a) except only top and bottom grade quartiles are included. Top-half scorers on the Math pretest had more high grades and fewer low grades than low-half scorers on that test. Ovals are drawn to enclose roughly the top and bottom 50% of Math scores, respectively. (The varying separations between Math scores are artifacts introduced by the conversion to within-class percentiles and the differing numbers of students in each of the three courses combined in this sample.)

**B. Association between pretest scores and course grades: An illustration**

As an illustration of the impact of the strong association between diagnostic pretest scores and course grades, a scatterplot of Lawson pretest scores and course grades for our largest sample (Alg-1 2005 CU, $N = 469$) is shown in Fig. 3. This figure is similar to Fig. 2 except that it uses Lawson scores instead of Math scores, and it uses actual grade points received on a 0-100% scale rather than class percentiles; this accounts for the more compact appearance of the grade distribution. The adjusted R-squared at only +0.156 is not particularly large and implies that most of the variance in the grades is due to other factors. Nonetheless, the median grade for those students who scored 63% on the Lawson test (15 out of 24 correct; $N = 43$) was only C (72.3 grade points), while that for students who scored 88% (21 out of 24; $N = 59$) was B (81.6 grade points), a full letter grade higher. This high degree of association between grades and pretest scores, though illustrated with particular clarity in this large sample, was quite typical of the great majority of the 116 cases examined. (We note that some results for this class have already been presented by Dubson and Pollock; see ref. 38.)

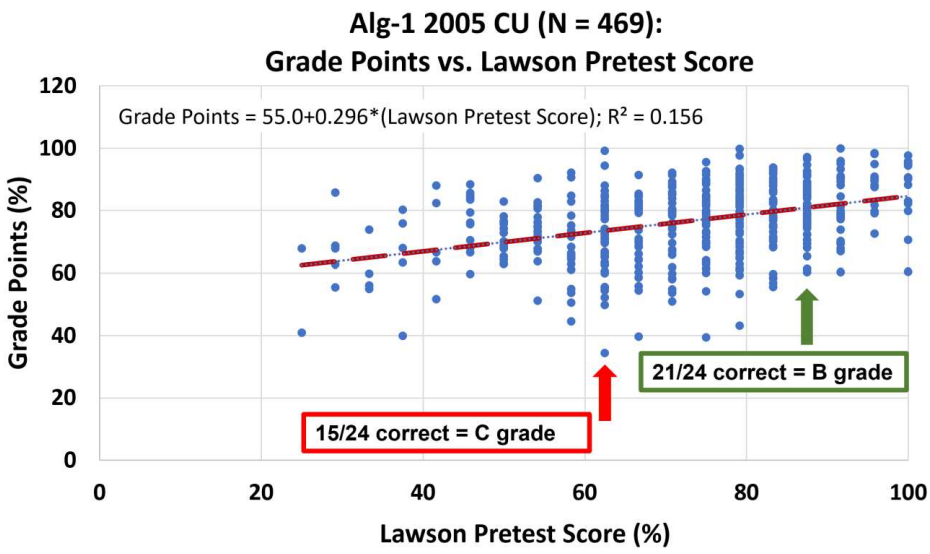



Fig. 3. Final course grade points (0-100% scale) for students in the Alg-1 2005 CU sample as a function of their score on the Lawson pretest (0-100% scale); each dot represents one student. The equation for the linear fit line is shown, as are the median letter grades for students scoring either 15 (63%) or 21 (88%) (out of 24) on the Lawson pretest. (Grade point cutoffs were 70% for a C and 80% for a B.)

Although it would be useful to assess the implications for students who score in the bottom quartile on two or more of the diagnostic tests ("bottom-bottom"), the small sample sizes make this impractical in most cases since relatively few students fit into this category. One exception is

the Alg-1 2005 CU sample, in which 30 students (out of 469 total) scored in the bottom quartile on *both* the FCI and the Lawson test. Of these 30 students, not a single one had a final course grade within the top quartile of the class. (For a random selection of 30 students, 7 or 8 would have been expected to score in the top quartile.) At the same time, their proportion of bottom-quartile grades was not exceptionally high. In the Alg-2 AAA sample that combined three separate courses (2022, 2023, and 2024), only two of the 23 students (9%) who scored in the bottom quartile on both the Math diagnostic and the Lawson test had final grades in the top quartile, compared to approximately six expected by random selection. Moreover, 10 (43%) of the students in this group had bottom-quartile final grades. For this sample, the proportions of both top- and bottom-quartile grades for the "bottom-bottom" group were thus somewhat worse than either the bottom-quartile Math or bottom-quartile Lawson groups on their own, although not dramatically so (see Fig. 1, caption).

### C. Relative predictive power of the three pretests

It is clearly of interest to determine which of the pretests, if any, might offer the greatest power in predicting grade probabilities. However, the very great differences between the nature and topics of the different courses and between the methods and procedures of different instructors, the varying grading philosophies, and differences among student populations at different institutions all contribute to making this type of comparison difficult indeed. Moreover, the generally small sample sizes exacerbate the problem. For the bulk of our samples, these challenges proved unresolvable. More detailed discussion can be found in the Supplementary Material, along with a table of all regression coefficients and associated *p*-values for multiple linear regression fits for all samples in which more than one pretest was administered.

### D. Correlations among the pretest scores

Certainly there are, as one might expect, significant positive correlations between scores on any one of the pretests and scores on the other(s) in classes in which two or more pretests were administered. For example, the average correlation coefficients between FCI and Lawson pretest scores were in the range 0.40-0.45 and between FCI and Math they were 0.30-0.35, while those between Math and Lawson were 0.30 and 0.45 in Alg-1 and Alg-2 respectively. Further discussion of the implications of these correlations can be found in the Supplementary Material.

### E. Analysis of largest samples: which pretest is most closely associated with final grade?

In order to explore in greater depth the relative degree of association with course grades of the various pretests, we carried out a very detailed analysis of our largest samples. For Alg-1 2005 CU, it seemed clear that Lawson pretests scores were more closely associated with grades than were FCI pretest scores. For example, a multiple regression calculation in this case yielded the equation $G = 54.611 + 0.264L + 0.081F$ where $G$, $L$, and $F$ are the percentile scores for final course grade, Lawson pretest score, and FCI pretest score, respectively. Both $L$ and $F$ were statistically significant predictors in this model; however, the much larger weight of the $L$ coefficient as well as other comparison criteria seemed to favor in more unambiguous fashion the Lawson pretest scores over FCI scores as the more influential of the two variables in this sample. Further details are provided in the Supplemental Material.

It may seem plausible that Lawson (reasoning) pretest scores would be a more influential factor in predicting final grades than FCI pretest scores. However, since we have only a single sample large enough to carry out this comparison, we hesitate to draw any broader conclusions from this finding. Depending on the previous level of physics preparation of a particular class, the relative weight of Lawson and FCI pretest scores could conceivably vary substantially.

For the combined sample Alg-2 ASU-P 2022-23-24, there was no clear-cut resolution in favor of either Math or Lawson scores. A multiple regression calculation in this case yielded the equation $G = 26.363 + 0.185L + 0.295M$, where $G$, $L$, and $M$ are the percentile scores for final course grade, Lawson pretest score, and Math pretest score, respectively; both $L$ and $M$ were statistically significant predictors in this model. Since the coefficient of $M$ is substantially larger than that of $L$, the implication might be that the Math pretest score was the more "influential" of the two pretests in this case. However, some of the other comparison criteria contradicted this conclusion. The details of these calculations may be found in the Supplementary Material.

**F. Interaction effects**

A so-called "interaction" effect in the present context might imply, for example, that grades are more strongly dependent on (e.g., have higher correlation with) Math pretest scores for students having high Lawson scores than they are for students with low Lawson scores. (The opposite pattern in which Lawson scores and Math scores exchange places would also be an interaction effect.) Although standard multiple regression techniques did not show any significant interaction effects, a closer analysis of our data suggested the possibility of an effect in which the magnitude of the correlation between grades and scores on one of the pretests is larger for students who score higher on the *other* pretest. One might characterize this relationship by saying that if a student scores low on one of the pretests, their score on the *other* pretest is less predictive of their grade than it might be if they had scored high on that pretest. However, this does not imply that high pretest scores on both or all pretests are *necessary* for success in the course. Rather, it suggests that for students who score high on one of the pretests, a high score on the other pretest is more *likely* to be accompanied by a high grade than would be the case if the score on the first pretest had been low. (Certainly, "high-high" scores are more likely to be associated with high grades than are "low-low" scores but scoring high on one pretest and just average on another is also associated with a high grade.) Further details are in the Supplementary Material.

**G. Comment on differences among the samples**

The sample includes classes from three state universities (two large, one medium) and a medium-size private research institution; some differences among the student populations would be expected. The only objective measures we have at hand to compare them are the pretest scores, but it would not make sense to compare, for example, pretest scores in an algebra-based course at one institution to those in a calculus-based course at another. Arguably the only course for which we have adequate numbers to compare is the first-semester algebra-based course, and so we note that in this course the average top-quartile cut-offs for Lawson score (maximum = 24) were approximately 21, 20, and 17 for courses at CU ($N = 1$), LMU ($N = 10$), and ASU-P ($N = 6$), respectively. The bottom-quartile cutoffs for the same test in the same sequence were 15, 14,

and 10. For the FCI (maximum score = 30), top-quartile cutoffs were 13, 12, and 11, while bottom-quartile cutoffs were 7, 5, and 5. (For LMU, $N = 9$ for FCI.) These figures indicate a significant difference between ASU-P and the other two universities, at least regarding Lawson test scores. We did not attempt to investigate any implications of these differences, but it is notable that the pattern of grade probabilities and odds ratios was quite consistent among the three institutions.

Other factors that could conceivably have resulted in differences among the samples would be whether they were algebra-based or calculus-based classes, whether they were first-semester (mechanics) or second-semester (E&M) classes, or whether they might have been affected by the Covid era in varying ways, including the dates of the courses. We did not observe any clear-cut signals of effects due to any of these other factors, since the variation in odds ratios was so much greater than any that might have been due to these factors. Of course, in the future, with a larger sample including larger classes, such differences might become apparent.

## IV. IMPLICATIONS AND CONCLUSIONS

Our findings that high scores on mathematics, reasoning, and physics concept pretests are all associated with higher course grades in introductory physics are not themselves new; rather, they are consistent with findings of many other investigations over the years. However, we offer a novel approach to analyzing this association by comparing explicitly the probabilities of obtaining high and low course grades by the high and low scorers on the pretests. We have pointed out that although linear correlation coefficients between grades and pretest scores are generally fairly low, the relative probabilities of obtaining high and low course grades by high and low pretest scorers are *very* different, generally differing by factors between 2–6. Apart from a 40-year-old report that addressed similar themes, it seems that only recently has the possibility of such a large difference again been pointed out, and that only based on model calculations by Salehi et al.; our investigation offers strong support for that model prediction. It is worth elaborating on this point in more detail.

Salehi et al. developed a number of predictive models for final exam performance in the introductory calculus-based mechanics course ("physics 1"); their models incorporated scores on pre-instruction tests of both math skills (in the form of the SAT or ACT) and physics concept knowledge (in the form of the FCI or the Force and Motion Conceptual Evaluation, another mechanics diagnostic test[42]). They found that typical values of adjusted $R$-squared for their model fits were in the range 0.20-0.30, roughly in agreement with our own findings, and they made the following crucial observation based on a model calculation that assumed normally distributed measures of incoming preparation:

> *These R-squared values may seem modest to some, but they have career-altering implications for students who are poorly prepared….for an R-squared of 0.34…a student who comes in with preparation in the bottom quartile has about a factor of 4 higher probability of being in the bottom quartile of the grade distribution than a student who starts the course in the upper quartile of preparation. If one considers bottom quartile exam scores as failing, this means that poorly prepared students are 4 times more likely to fail their physics 1 final exams than peers with good incoming preparation. [Salehi et al. (2019), p. 020114-6]*

Although one might argue that merely obtaining a bottom-quartile grade is not in itself necessarily associated with "failing," their quantitative prediction is in perfect agreement with our empirical finding that bottom-quartile scorers on pretests are 3-5 times more likely to obtain bottom-quartile grades than peers who scored in the top quartile on the pretests. Moreover, their point regarding the potential career-altering implications for poorly prepared students enrolled in this crucial "gateway" course is well-taken and one with which we fully concur.

It is essential to underline that none of these studies, including our own, have empirically tested whether there is a *causal* relationship between physics grades and the skills measured by the various pretests. (Such a test is distinct from the creation of so-called statistical "causal models.") We would argue that there is no firm basis for asserting, for example, that efforts to improve students' pre-instruction physics concept knowledge, or math and reasoning skills, would lead directly to improvements in these students' physics course performance. However plausible such an expectation might be, it is nonetheless conceivable that these various pretest measures are merely markers associated with other aspects of students' pre-college experiences that have more direct impact on their ability to succeed in college physics. That said, we would agree that efforts to improve students' pre-college preparation are extremely desirable and offer perhaps the best prospects currently available for having significant impact on students' performance in college physics courses.

Needless to say, there are many other factors—unexplored here—that can impact a physics student's course performance. Arguably one of the most important is "motivation," the degree to which the student puts time, effort, and commitment into their physics study. Frequently mentioned by previous researchers, motivation has proven difficult to investigate systematically. Although we were unable to look at motivation in any systematic way in our own investigation, one of the instructors had many opportunities to study the students in his own classes that formed part of the data sample. It was remarkable how often the "outlier" cases could plausibly be explained by motivational factors. Students who scored very high on pretests but ended with very low course grades frequently missed classes, failed to turn in assignments, and/or did not participate in class problem-solving activities. By contrast, students who scored low on the pretests but achieved high course grades typically attended class consistently, participated actively in class activities and frequently asked questions, turned in assignments on time, and often sought additional aid from the instructor via e-mail, Zoom, and/or after-class review sessions. Additional light on the "motivation" issue might be shed by examining students' performance on course exams and comparing those performances to their final course grades. Elements of course grades such as attendance, homework, and other assignments are often associated with motivational factors and can indicate levels of engagement. However, for the great majority of our samples, we did not collect the data that would be needed for this type of analysis and so it will have to remain for future investigators to examine.

We did not track and therefore were unable to consider a number of other pre-instruction factors that have featured prominently in recent investigations, including among others gender, demographic variables, self-efficacy, high-school background, and attitudes about learning science. Another recent study has looked at whether the nature of the instructional methods impacts the apparent influence of some of the pre-instruction factors.[43] As mentioned previously, the instruction in most of the courses in our sample incorporated a variety of research-based,

active-learning approaches; this did not really offer an opportunity to examine the nature of instruction as a possible confounding factor.

It is, of course, of great interest and potential importance to understand why, exactly, the predictive power of the various pretests is so large, representing—as they do—the previous preparation of the students enrolled in introductory physics classes. Perhaps the most important point to make is one that can sometimes be overlooked by college physics instructors, that is: Before we meet the students in our classes, they have 16-18 years of previous experience, typically including 13 years or more of formal academic preparation. That experience includes exposure to mathematics, to at least some physics, and to many "reasoning"-related tasks. It would be rather strange if those many years of preparation did not have some significant impact on the potential success of those students in our classes, regardless of the great effort and diverse instructional methods we bring to bear as instructors.

We have some additional clues. In unpublished work, we have found that class-average score on most individual items on the mathematics pretest is *highly* predictive of that class's average score on the full 16-item test, even though many different topics are included on the test. (The class-average item score is simply the proportion of all students in a class who got that item correct.) There is a very wide spread in class preparation, as indicated by class-average scores on the full diagnostic ranging from 62% to 92% for 27 different classes at multiple institutions. However, merely knowing the class-average score on any one of about half of the individual test items allows a prediction of the class-average score on the full diagnostic to ±5% with roughly 90% confidence. This suggests that difficulties with math skills are not centered on one or another such skill but instead tend to cluster together. This indicates that Math pretest scores represent a broad category of pre-college preparation that is reflected in a multitude of individual skills, not easily addressed by one or two college math courses. Similarly, whatever student abilities are measured by the Lawson reasoning test, they have been under development for well over a decade and are not likely to change significantly during a single four-month semester. The dynamics of how these reasoning skills may interact with physics instruction were subjects of investigation during the 1970s by Karplus, Arons, and others; these authors emphasized that an ability—and inclination—to search for relationships and patterns was central to success in physics.[44]

In spite of the many factors that influence students' physics course performance, our investigation has one clear and consistent outcome that bears repeating, that is: the high- and low-grade probabilities of high-scorers on a single pre-instruction diagnostic test typically differ by a factor of three to six from the corresponding grade probabilities of low scorers on that pretest. Since we are referring here only to probabilities, the course performance outcome for any *individual* student, regardless of their pretest scores, remains highly uncertain. However, it is reasonable to acknowledge that the course performance expected for the broad *group* of low-scorers on these pretests must be very different from that expected for the group of high-scorers.

A natural follow-up question is whether this finding might be used in some way to offer modified or supplemental instruction for the more "at-risk" group that could perhaps improve their course outcomes. We should emphasize that most of the instructors of the courses in our samples were already both highly motivated and very experienced in providing activity-based instruction that engages students much more intensively than traditional "lecture only" formats. They used methods and materials that were based on or informed by physics education research,

and which have proven track records of improving student outcomes. Among the explicit goals and demonstrated outcomes of these methods is to enhance the learning of students who have substandard preparation (see Ref. 40 and references therein). This suggests that improving the situation in ways that might, for example, reduce the outcome gap between high- and low-pretest scorers will not be easy. In principle, instruction that is modified and more precisely targeted at the most at-risk students might show improved effectiveness, but implementation of such measures is associated with a host of practical and logistical challenges.

Previous investigators have raised similar questions and offered suggestions such as math skills enrichment, practice on reasoning skills, and modifications of the instructional methods. Some reports have suggested promising results when implementing these measures (most recently Ref. 43); however, none appear to have attempted to assess the effectiveness of such measures by explicitly examining changes in high- and low-grade probabilities. There is no question that an investigation of this type would be fraught with methodological challenges. Nonetheless, there are potential advantages to using a relatively straightforward high/low grade comparison as an assessment measure in contrast to attempting to create and test complicated statistical models, as others have done. We hope that further research will shed additional light on methods for improving the course-performance outlook of students who enter with significantly weaker preparation as measured by pre-instruction diagnostic tests.

## SUPPLEMENTARY MATERIAL

Detailed discussions and data analysis related to Sections III.C, D, E, and F are included in the Supplementary Material. A print version of the mathematics diagnostic test is also included; for access to the automated online version, please contact the corresponding author.

## ACKNOWLEDGMENTS

We are very grateful to our colleagues who provided data for this investigation, including Vince Coletta, Chris Varney, Steve Pollock, and Rizal Hariadi. John Byrd was instrumental in initiating the grade-analysis efforts. We thank Valerie Otero for suggesting that we investigate low-grade odds ratios, and we also had many valuable discussions with Andrew Heckler. This work was supported in part by NSF DUE #1504986 and #1914712.

## AUTHOR DECLARATIONS

### Conflict of Interest

The authors have no conflicts to disclose.

**Author Contributions**

Both authors designed the study and the method of data analysis. Dakota King developed the data analysis code and carried out the bulk of the initial analysis. David Meltzer completed the analysis and wrote the paper; Dakota King checked the manuscript and approved it.

**Data Availability**

The data that support the findings of this study are available from the corresponding author upon reasonable request

## APPENDIX

Table AI. High-grade probability vs. Math pretest score. The columns show the percentages of students with top-quartile scores ("Top-quartile Math") and bottom-quartile scores ("Bottom-quartile Math") on the mathematics pretest who received top-quartile grades in their class. The high-grade odds ratio is found by dividing the value of the grade percentage of the top-quartile group by that of the bottom-quartile group. The bottom row shows total *N* [in brackets] and unweighted averages of the top- and bottom-quartile columns, while the ratio of those two averages is shown in the Odds Ratio column. (Course and campus code in Table 1.) An outlier case (that does not follow the general pattern) is shown in bold red italics.

| Course | Campus | *N* | Top-quartile Math | Bottom-quartile Math | High-grade odds ratio |
|---|---|---|---|---|---|
| Alg-1 2021a | ASU-P | 39 | 51% | 10% | 5.0 |
| Alg-1 2021b | ASU-P | 42 | 44% | 10% | 4.6 |
| Alg-1 2022a | ASU-P | 40 | 27% | 6% | 4.4 |
| Alg-1 2022b | ASU-P | 52 | 49% | 10% | 5.1 |
| Alg-1 2023a | ASU-P | 42 | 39% | 10% | 4.1 |
| Alg-1 2023b | ASU-P | 46 | 64% | 9% | 7.3 |
| Alg-2 2022 | ASU-P | 75 | 46% | 21% | 2.2 |
| Alg-2 2023 | ASU-P | 92 | 41% | 13% | 3.2 |
| Alg-2 2024 | ASU-P | 99 | 51% | 8% | 6.3 |
| ***Alg-2 2021*** | ***ASU-T*** | ***129*** | ***30%*** | ***39%*** | ***0.8*** |
| Calc-1 2021a | UWF | 53 | 38% | 0% | [undefined] |
| Calc-1 2021b | UWF | 42 | 44% | 0% | [undefined] |
| Calc-2 2021 | UWF | 58 | 43% | 14% | 3.1 |
| **AVERAGE (unweighted)** | | **809** | **44%** | **12%** | **3.8** |

Table AII. Low-grade probability vs. Math pretest score. The columns show the percentages of students with top-quartile scores ("Top-quartile Math") and bottom-quartile scores ("Bottom-quartile Math") on the mathematics pretest who received bottom-quartile grades in their class. The Low-grade odds ratio is found by dividing the value of the grade percentage of the bottom-quartile group by that of the top-quartile group. The bottom row shows total *N* [in brackets] and unweighted averages of the top- and bottom-quartile columns, while the ratio of those two averages (bottom divided by top) is shown in the Odds Ratio column. (Course and campus code in Table 1.)

| Course | Campus | *N* | Top-quartile Math | Bottom-quartile Math | Low-grade odds ratio |
|---|---|---|---|---|---|
| Alg-1 2021a | ASU-P | 39 | 10% | 41% | 4.0 |
| Alg-1 2021b | ASU-P | 42 | 16% | 48% | 3.0 |
| Alg-1 2022a | ASU-P | 40 | 0% | 42% | [undefined] |
| Alg-1 2022b | ASU-P | 52 | 26% | 29% | 1.1 |
| Alg-1 2023a | ASU-P | 42 | 20% | 31% | 1.5 |
| Alg-1 2023b | ASU-P | 46 | 3% | 21% | 7.3 |
| Alg-2 2022 | ASU-P | 75 | 11% | 26% | 2.4 |
| Alg-2 2023 | ASU-P | 92 | 11% | 30% | 2.8 |
| Alg-2 2024 | ASU-P | 99 | 5% | 45% | 8.4 |
| Alg-2 2021 | ASU-T | 129 | 11% | 30% | 2.8 |
| Calc-1 2021a | UWF | 53 | 0% | 47% | [undefined] |
| Calc-1 2021b | UWF | 42 | 14% | 38% | 2.8 |
| Calc-2 2021 | UWF | 58 | 24% | 44% | 1.8 |
| **AVERAGE (unweighted)** | | **809** | **12%** | **36%** | **3.1** |

| Course | Campus | *N* | Top-quartile FCI | Bottom-quartile FCI | High-grade odds ratio |
|---|---|---|---|---|---|
| Alg-1 2017 | ASU-P | 22 | 55% | 0% | [undefined] |
| Alg-1 2018 | ASU-P | 53 | 45% | 8% | 6.0 |
| Alg-1 2019 | ASU-P | 63 | 38% | 13% | 3.0 |
| Alg-1 2021a | ASU-P | 35 | 57% | 0% | [undefined] |
| Alg-1 2021b | ASU-P | 37 | 32% | 17% | 1.9 |
| Alg-1 2022a | ASU-P | 41 | 21% | 15% | 1.4 |
| Alg-1 2022b | ASU-P | 52 | 26% | 7% | 3.9 |
| Alg-1 2023a | ASU-P | 40 | 30% | 20% | 1.3 |
| Alg-1 2023b | ASU-P | 47 | 55% | 18% | 3.1 |
| Alg-1 2005 | CU | 470 | 41% | 12% | 3.5 |
| Alg-1 2007 | LMU | 23 | 87% | 0% | [undefined] |
| Alg-1 2009 | LMU | 51 | 63% | 0% | [undefined] |
| Alg-1 2012 | LMU | 44 | 50% | 0% | [undefined] |
| Alg-1 2013 | LMU | 30 | 51% | 0% | [undefined] |
| Alg-1 2014 | LMU | 33 | 43% | 12% | 3.6 |
| Alg-1 2015 | LMU | 24 | 67% | 0% | [undefined] |
| Alg-1 2016 | LMU | 34 | 71% | 0% | [undefined] |
| Alg-1 2018 | LMU | 47 | 34% | 14% | 2.4 |
| Alg-1 2021 | LMU | 27 | 44% | 0% | [undefined] |
| Calc-1 2012 | ASU-P | 40 | 43% | 0% | [undefined] |
| Calc-1 2013a | ASU-P | 18 | 44% | 0% | [undefined] |
| Calc-1 2013b | ASU-P | 48 | 54% | 17% | 3.3 |
| Calc-1 2021a | UWF | 62 | 29% | 26% | 1.1 |
| Calc-1 2021b | UWF | 53 | 40% | 15% | 2.6 |
| **AVERAGE (unweighted)** | | **(1394)** | **47%** | **8%** | **5.8** |

Table AIII. High-grade probability vs. FCI pretest score. The columns show the percentages of students with top-quartile scores ("Top-quartile FCI") and bottom-quartile scores ("Bottom-quartile FCI") on the FCI pretest who received top-quartile grades in their class. The high-grade odds ratio is found by dividing the value of the grade percentage of the top-quartile group by that of the bottom-quartile group. The bottom row shows total *N* [in brackets] and unweighted averages of the top- and bottom-quartile columns, while the ratio of those two averages is shown in the Odds Ratio column. (Course and campus code in Table 1.)

| Course | Campus | *N* | Top-quartile FCI | Bottom-quartile FCI | Low-grade odds ratio |
|---|---|---|---|---|---|
| Alg-1 2017 | ASU-P | 22 | 18% | 32% | 1.8 |
| Alg-1 2018 | ASU-P | 53 | 19% | 45% | 2.4 |
| Alg-1 2019 | ASU-P | 63 | 6% | 47% | 7.4 |
| Alg-1 2021a | ASU-P | 35 | 0% | 56% | [undefined] |
| Alg-1 2021b | ASU-P | 37 | 11% | 43% | 4.0 |
| Alg-1 2022a | ASU-P | 41 | 21% | 39% | 1.9 |
| Alg-1 2022b | ASU-P | 52 | 18% | 33% | 1.8 |
| Alg-1 2023a | ASU-P | 40 | 20% | 37% | 1.8 |
| Alg-1 2023b | ASU-P | 47 | 9% | 43% | 5.1 |
| Alg-1 2005 | CU | 470 | 19% | 22% | 1.1 |
| Alg-1 2007 | LMU | 23 | 0% | 52% | [undefined] |
| Alg-1 2009 | LMU | 51 | 8% | 47% | 6.0 |
| Alg-1 2012 | LMU | 44 | 9% | 50% | 5.4 |
| Alg-1 2013 | LMU | 30 | 24% | 37% | 1.5 |
| Alg-1 2014 | LMU | 33 | 7% | 32% | 4.7 |
| Alg-1 2015 | LMU | 24 | 0% | 67% | [undefined] |
| Alg-1 2016 | LMU | 34 | 12% | 47% | 4.0 |
| Alg-1 2018 | LMU | 47 | 15% | 31% | 2.2 |
| Alg-1 2021 | LMU | 27 | 0% | 44% | [undefined] |
| Calc-1 2012 | ASU-P | 40 | 10% | 43% | 4.3 |
| Calc-1 2013a | ASU-P | 18 | 0% | 44% | [undefined] |
| ***Calc-1 2013b*** | ***ASU-P*** | ***48*** | ***17%*** | ***8%*** | ***0.5*** |
| Calc-1 2021a | UWF | 62 | 13% | 40% | 3.1 |
| Calc-1 2021b | UWF | 53 | 8% | 25% | 3.3 |
| **AVERAGE (unweighted)** | | ***N*=1394** | **11%** | **40%** | **3.7** |

Table AIV. Low-grade probability vs. FCI pretest score. The columns show the percentages of students with top-quartile scores ("Top-quartile FCI") and bottom-quartile scores ("Bottom-quartile FCI") on the FCI pretest who received bottom-quartile grades in their class. The Low-grade odds ratio is found by dividing the value of the grade percentage of the bottom-quartile group by that of the top-quartile group. The bottom row shows total *N* [in brackets] and unweighted averages of the top- and bottom-quartile columns, while the ratio of those two averages (bottom divided by top) is shown in the Odds Ratio column. (Course and campus code in Table 1. An outlier case (that does not follow the general pattern) is shown in bold red italics.)

[1] Archer Willis Hurd, *Problems of Science Teaching at the College Level* (University of Minnesota Press, Minneapolis, 1929), Parts III-V.

[2] Haym Kruglak and Robert J. Keller. "The prediction of achievement in sophomore engineering physics at the University of Minnesota," Am. J. Phys. **18**, 140-146 (1950).

[3] Ralph H. Blumenthal, "Split Sections and Learning in College Physics." Am. J. Phys. **25**, 352-355 (1957).

[4] Ralph H. Blumenthal, ''Multiple instruction and other factors related to achievement in college physics,'' Sci. Educ. **45**, 336–342 (1961).

[5] John R. Bolte, ''Background factors and success in college physics,'' J. Res. Sci. Teach. **4**, 74–78 (1966).

[6] George Barnes, ''Scores on a Piaget-type questionnaire versus semester grades for lower-division college physics students,'' Am. J. Phys. **45**, 841–847 (1977).

[7] Daniel Cohen, Donald F. Hillman, and Russell M. Agne, ''Cognitive level and college physics achievement,'' Am. J. Phys. **46**, 1026–1029 (1978).

[8] Dov Liberman and H. T. Hudson, ''Correlation between logical abilities and success in physics,'' Am. J. Phys. **47**, 784–786 (1979).

[9] Morris A. Enyeart, Dale Baker, and Dave Vanharlingen, ''Correlation of inductive and deductive logical reasoning to college physics achievement,'' J. Res. Sci. Teach. **17**, 263–267 (1980).

[10] Audrey B. Champagne, Leopold E. Klopfer, and John H. Anderson, ''Factors influencing the learning of classical mechanics,'' Am. J. Phys. **48**, 1074–1079 (1980).

[11] Audrey B. Champagne and Leopold E. Klopfer, ''A causal model of students' achievement in a college physics course,'' J. Res. Sci. Teach. **19**, 299–309 (1982).

[12] Warren Wollman and Frances Lawrenz, ''Identifying potential 'dropouts' from college physics classes,'' J. Res. Sci. Teach. **21**, 385–390 (1984).

[13] W. T. Griffith, ''Factors affecting performance in introductory physics courses,'' Am. J. Phys. **53**, 839–842 (1985).

[14] Ibrahim Abou Halloun and David Hestenes, ''The initial knowledge state of college physics students,'' Am. J. Phys. **53**, 1043–1055 (1985).

[15] H. T. Hudson, ''A comparison of cognitive skills between completes and dropouts in a college physics course,'' J. Res. Sci. Teach. **23**, 41–50 (1986).

[16] Susan McCammon, Jeannie Golden, and Karl K. Wuensch, ''Predicting course performance in freshman and sophomore physics courses: women are more predictable than men,'' J. Res. Sci. Teach. **25**, 501–510 (1988).

[17] H. T. Hudson and W. R. McIntire, ''Correlation between mathematical skills and success in physics,'' Am. J. Phys. **45**, 470–471 (1977).

[18] H. T. Hudson and Ray M. Rottmann, ''Correlation between performance in physics and prior mathematics knowledge,'' J. Res. Sci. Teach. **18**, 291–294 (1981).

[19] H. T. Hudson and Dov Liberman, ''The combined effect of mathematics skills and formal operational reasoning on student performance in the general physics course,'' Am. J. Phys. **50**, 1117–1119 (1982).

[20] David Hestenes, Malcolm Wells, and Gregg Swackhamer, "Force concept inventory," Phys. Teach. **30**(3), 141-158 (1992).

[21] Shima Salehi, Eric Burkholder, G. Peter Lepage, Steven Pollock, and Carl 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**(2), 020114 (2019).

[22] John Stewart, Geraldine L. Cochran, Rachel Henderson, Cabot Zabriskie, Seth DeVore, Paul Miller, Gay Stewart, and Lynnette Michaluk, "Mediational effect of prior preparation on performance differences of students underrepresented in physics," Phys. Rev. Phys. Educ. Res. **17**, 010107 (2021).

[23] Rachel Henderson, Dona Hewagallage, Jake Follmer, Lynnette Michaluk, Jessica Deshler, Edgar Fuller, and John Stewart, "Mediating role of personality in the relation of gender to self-efficacy in physics and mathematics," Phys. Rev. Phys. Educ. Res. **18**(1), 010143 (2022).

[24] Gerald E. Hart and Paul D. Cottle, ''Academic backgrounds and achievement in college physics,'' Phys. Teach. **31**, 470–475 (1993).

[25] Brian J. Alters, ''Counseling physics students: A research basis,'' Phys. Teach. **33**, 413–415 (1995).

[26] Philip M. Sadler and Robert H. Tai, ''Success in introductory college physics: The role of high school preparation,'' Sci. Educ. **85**, 111–136 (2001).

[27] Z. Hazari, R. H. Tai, and P. M. Sadler, "Gender differences in introductory university physics performance: The influence of high school physics preparation and affective factors," Sci. Educ. **91**, 847-876 (2007).

[28] M. Greene and R. Lopez, "Evaluating a predictive model of student performance in introductory calculus-based physics," IOP Conf. Series: Journal of Physics: Conf. Series **1286**, 012021 (2019)

[29] Alysa Malespina, Christian D. Schunn, and Chandralekha Singh, "Whose ability and growth matter? Gender, mindset and performance in physics," Int. J. STEM Educ. **9**(1), 28 (2022).

[30] Yangqiuting Li and Chandralekha Singh, "Sense of belonging is an important predictor of introductory physics students' academic performance," Phys. Rev. Phys. Educ. Res. **19**(2), 020137 (2023).

[31] David E. Meltzer, "The relationship between mathematics preparation and conceptual learning gains in physics: A possible 'hidden variable' in diagnostic pretest scores," Am. J. Phys. **70**, 1259-1268 (2002).

[32] Lin Ding, "Verification of causal influences of reasoning skills and epistemology on physics conceptual learning," Phys. Rev. Phys. Educ. Res. **10**(2), 023101 (2014).

[33] Vincent P. Coletta and Jeffrey A. Phillips, "Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability." Am. J. Phys. **73**, 1172-1182 (2005).

[34] Vincent P. Coletta, Jeffrey A. Phillips, and Jeffrey J. Steinert, "Interpreting force concept inventory scores: Normalized gain and SAT scores," Phys. Rev. Phys. Educ. Res. **3**(1), 010106 (2007).

[35] Vincent P. Coletta and Jeffrey J. Steinert, "Why normalized gain should continue to be used in analyzing preinstruction and postinstruction scores on concept inventories," Phys. Rev. Phys. Educ. Res. **16**(1), 010108 (2020).

[36] Vincent P. Coletta, "Evidence for a normal distribution of normalized gains," Phys. Rev. Phys. Educ. Res. **19**(1), 010111 (2023).

[37] Anton E. Lawson, "The development and validation of a classroom test of formal reasoning," J. Res. Sci. Teach. **15**, 11-24 (1978).

[38] M. A. Dubson and S. J. Pollock, "Can the Lawson test predict student grades?," AAPT Announcer **36**, 90 (2006).

[39] Steven J. Pollock, "Comparing student learning with multiple research-based conceptual surveys: CSEM and BEMA," In *AIP Conference Proceedings*, vol. 1064, no. 1, pp. 171-174 (American Institute of Physics, 2008).

[40] David E. Meltzer and Ronald K. Thornton, "Resource Letter ALIP-1: Active-Learning Instruction in Physics," Am. J. Phys. **80**, 478-496 (2012). The general nature of the methods used by these instructors is described in detail in Section V of this reference.

[41] https://www.physport.org/assessments/assessment.cfm?A=CTSR&S=4

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

[43] David J. Webb and Cassandra A. Paul, "Attributing equity gaps to course structure in introductory physics," Phys. Rev. Phys. Educ. Res. **19**(2), 020126 (2023).

[44] Some key references are the following: *Workshop on Physics Teaching and the Development of Reasoning: Complete Set of Modules*, edited by ; F. P. Collea, R. Fuller, R. Karplus, L. G. Paldy, and J. W. Renner (AAPT, Stony Brook, NY, 1975), https://digitalcommons.unl.edu/karplusworkshop/13/; A. B. Arons and R. Karplus, "Implications of accumulating data on levels of intellectual development," Am. J. Phys. **44**, 396 (1976); A. B. Arons, "Cultivating the capacity for formal reasoning: Objectives and procedures in an introductory physical science course," Am. J. Phys. **44**, 834-838 (1976); *College Teaching and the Development of Reasoning*, edited by R. G. Fuller, T. C. Campbell, D. I. Dykstra, Jr., and S. M. Stevens (Information Age Publishing, Charlotte, NC, 2009).

# Pre-instruction diagnostic tests can predict grade probabilities in introductory physics

## Supplementary Material

David E. Meltzer [a)]

School of Applied Sciences and Arts

College of Integrative Sciences and Arts

Arizona State University

Mesa, AZ 85212

Dakota H. King

University of Arizona College of Veterinary Medicine

Oro Valley, Arizona 85737

a) Corresponding author; david.meltzer@asu.edu

## II D: Determination of grade probabilities and odds ratios

This figure illustrates the example provided in Section II D.

| Student Number | Score on Math Pretest (number correct out of 16) | Group designation | Grade (percentile) |
|---|---|---|---|
| 1 | 16 | F | 81 |
| 2 | 16 | F | 95 |
| 3 | 15 | F | 60 |
| 4 | 15 | F | 40 |
| 5 | 15 | F | 98 |
| 6 | 14 | F | 67 |
| 7 | 14 | F | 83 |
| 8 | 14 | F | 36 |
| | | | |
| 9 | 13 | S | 79 |
| 10 | 13 | S | 17 |
| 11 | 13 | S | 52 |
| 12 | 13 | S | 71 |
| 13 | 13 | S | 100 |
| | | | |
| 14 | 12 | | 43 |
| 15 | 12 | | 7 |
| 16 | 12 | | 93 |
| 17 | 12 | | 69 |
| 18 | 12 | | 24 |
| 19 | 12 | | 48 |
| 20 | 11 | | 45 |
| 21 | 11 | | 14 |
| 22 | 11 | | 74 |
| 23 | 11 | | 10 |
| 24 | 11 | | 33 |
| 25 | 10 | | 26 |
| 26 | 10 | | 57 |
| 27 | 10 | | 86 |
| 28 | 9 | | 38 |
| 29 | 9 | | 64 |
| 30 | 9 | | 21 |
| 31 | 9 | | 55 |
| 32 | 9 | | 90 |
| 33 | 8 | | 88 |
| 34 | 8 | | 50 |
| 35 | 8 | | 5 |
| 36 | 8 | | 62 |
| 37 | 7 | | 19 |
| 38 | 7 | | 31 |
| 39 | 6 | | 12 |
| 40 | 6 | | 2 |
| 41 | 5 | | 76 |
| 42 | 5 | | 29 |

**N** = 42
**Q** = 10.5 (Full sample, top quartile)
**F** = 8 (number of students whose math pretest scores were equaled or exceeded by no more than 10 students)
**S** = 5 (number of students whose math pretest scores were one point lower than the lowest in the F group)
**f** = (10.5-8)/5 = 0.5 = fraction of all S-group students needed to add to F group students such that F + (f*S) = Q
**4** = number of students with top-quartile (>75th percentile) grades in F group
**2** = number of students with top-quartile (>75th percentile) grades in S group
Number of students with top-quartile (>75th percentile) grades in top-quartile math group = 4 + (0.5)*2 = 5
Probability for top-quartile math scorers to obtain top-quartile grade = 5/10.5 = 48%

Fig. S0. An illustration of the example provided in Section II D of the main text, using simulated data. Red font indicates upper-quartile grades.

**III C: Relative predictive power of the three pretests**

It is clearly of interest to determine which of the pretests, if any, might offer the greatest power in predicting grade probabilities. However, the very great differences between the nature and topics of the different courses and between the methods and procedures of different instructors, the varying grading philosophies, and differences among student populations at different institutions all contribute to making this type of comparison difficult indeed. Moreover, the generally small sample sizes exacerbate the problem. One straightforward approach is simply to compare the odds ratios of the different pretests: does one or another of the pretests seem to generate a significantly larger grade probability ratio than the others? The average high-grade odds ratios are 5.8, 3.8, and 5.8 for Lawson, Math, and FCI pretests respectively. For the low-grade odds ratios, the corresponding values are 4.5, 3.1, and 3.7. These values do not differ greatly from each other and, given the enormous range and variability of the odds ratios in the various courses, it would be hard to claim from this with any confidence that, for example, the Lawson pretest seems more predictive than the Math pretest. Indeed, even within the same course at the same institution over a period of several years, there is a great deal of variation in the odds ratios of any of the individual pretests. (Stewart et al. [Ref. 22]) emphasize that various types of pre-instruction preparation may—and usually do—appear to influence grades in ways that differ significantly among different demographic groups. Still, the variation we observed suggests that additional factors beyond simple demographics are also at work.)

In those classes in which two or more of the pretests were administered, it is possible—in principle—to try to compare the relative predictive power of the different pretests. However, again, the generally small sample sizes make this extremely difficult. A straightforward approach is to carry out a multiple linear regression analysis for each class such that course grade is expressed as a function of the different pretest scores, and for which the relative magnitude and significance of the coefficients of the different pretests in the resulting predictor equation are compared. The results of such an analysis are that, in most cases, the coefficients of one or more of the individual pretests are found to be non-significant ($p>0.05$), even though the quartile odds ratios for these pretests are often quite large. To illustrate this observation, we can look at the results for Alg-1 2022b ASU-P ($N = 46$); the equation that is generated by a multiple linear regression is $Grade = 70.9 + 0.130\ Lawson + 0.037\ Math + 0.126\ FCI$ where *Grade*, *Lawson*, *Math*, and *FCI* refer to the course grade points and the pretest scores for the Lawson, Math, and FCI tests, all expressed on a 0-100% scale. The adjusted $R^2$ for the overall fit is 0.12 and the *p*-value is 0.04, indicating a fairly low but marginally significant correlation. (That is, the multi-factor predictor equation yields grades that are significantly closer to the actual observed values than would be the case if one simply assigned the average course grade to each student in the class.) However, the *p*-values for the individual coefficients are all greater than 0.05 ($p = 0.11$, 0.69, and 0.18 for Lawson, Math, and FCI respectively), none of which rises to the level of significance. Despite this seeming implication of low association between pretest scores and grades, Tables II and III and V-VIII show that high scorers on the pretests were far more likely (≈400-600%) to get high grades in this course and (with one exception) far less likely (≈35-55%) to get low grades, than low scorers in the same course. This pattern is representative of the other courses in the sample and is a caution that standard linear regression analysis may be highly misleading in similar cases.

More broadly, we have six samples in which all three diagnostic pretests were administered, all in Alg-1 at ASU-P, including 2021a,b; 2022a,b; and 2023a,b. In the multiple linear regression including all three pretests, the coefficients of predictor variables varied widely: 0.08-0.30 for Lawson; 0.10-0.37 for Math; -0.23-+0.26 for FCI. In none of the six samples did all three predictor variables meet the $p<0.05$ criterion for statistical significance. The Lawson coefficient met that criterion in two of the samples, as did the Math—but in only one of those samples did both the Lawson and Math coefficients meet the criterion. The FCI coefficient met the $p<0.05$ criterion in only one of the six samples, while neither the Lawson nor the Math coefficient met the criterion in that sample.

Table S0 shows all multiple correlation coefficients and their respective *p*-values for all courses in which more than one diagnostic was administered.

| Sample | *N* | Adjusted *R* -squared | Regression *p* -value | Lawson coefficient | Lawson *p* -value | Math coefficient | Math *p* -value | FCI coefficient | FCI *p* -value |
|---|---|---|---|---|---|---|---|---|---|
| Alg-1 2021a ASU-P | 35 | 0.30 | 0.00 | 0.08 | 0.44 | 0.10 | 0.32 | 0.26 | 0.02 |
| Alg-1 2021b ASU-P | 34 | 0.28 | 0.00 | 0.08 | 0.42 | 0.37 | 0.01 | 0.13 | 0.39 |
| Alg-1 2022a ASU-P | 38 | 0.28 | 0.00 | 0.17 | 0.04 | 0.34 | 0.00 | -0.23 | 0.08 |
| Alg-1 2022b ASU-P | 46 | 0.12 | 0.04 | 0.13 | 0.11 | 0.04 | 0.69 | 0.13 | 0.18 |
| Alg-1 2023a ASU-P | 34 | -0.03 | 0.55 | 0.12 | 0.49 | 0.10 | 0.47 | 0.01 | 0.97 |
| Alg-1 2023b ASU-P | 38 | 0.38 | 0.00 | 0.30 | 0.01 | 0.13 | 0.28 | 0.25 | 0.17 |
| Alg-1 2005 CU | 466 | 0.17 | 0.00 | 0.26 | 0.00 | | | 0.08 | 0.02 |
| Alg-1 2007 LMU | 24 | 0.72 | 0.00 | 0.27 | 0.00 | | | 0.15 | 0.07 |
| Alg-1 2009 LMU | 51 | 0.35 | 0.00 | 0.20 | 0.00 | | | 0.34 | 0.00 |
| Alg-1 2012 LMU | 44 | 0.36 | 0.00 | 0.20 | 0.01 | | | 0.25 | 0.02 |
| Alg-1 2013 LMU | 30 | 0.11 | 0.07 | 0.07 | 0.51 | | | 0.39 | 0.06 |
| Alg-1 2014 LMU | 33 | 0.50 | 0.00 | 0.42 | 0.00 | | | 0.07 | 0.44 |
| Alg-1 2015 LMU | 24 | 0.51 | 0.00 | 0.44 | 0.00 | | | 0.12 | 0.32 |
| Alg-1 2016 LMU | 34 | 0.51 | 0.00 | 0.27 | 0.00 | | | 0.26 | 0.01 |
| Alg-1 2018 LMU | 47 | 0.25 | 0.00 | 0.15 | 0.03 | | | 0.17 | 0.01 |
| Alg-1 2021 LMU | 27 | 0.60 | 0.00 | 0.35 | 0.00 | | | 0.13 | 0.13 |
| Calc-1 2021a UWF | 53 | 0.26 | 0.00 | | | 0.57 | 0.00 | -0.22 | 0.05 |
| Calc-1 2021b UWF | 41 | -0.01 | 0.43 | | | 0.07 | 0.62 | 0.12 | 0.32 |
| Alg-2 2022 ASU-P | 63 | 0.18 | 0.00 | 0.27 | 0.00 | 0.12 | 0.19 | | |
| Alg-2 2023 ASU-P | 67 | 0.03 | 0.15 | 0.10 | 0.12 | 0.03 | 0.70 | | |
| Alg-2 2024 ASU-P | 86 | 0.24 | 0.00 | 0.05 | 0.31 | 0.29 | 0.00 | | |

Table S0. For each course in which more than one diagnostic was administered, the adjusted *R*-squared of the multiple linear regression is shown along with its *p*-value. (Red font indicates *p* values meeting the $p<0.05$ significance criterion.) The regression coefficients of the Lawson, Math, and FCI variables are shown as well, along with their respective *p*-values. Although 17 of the 21 regressions are statistically significant, one or more of the predictor variable coefficients fails to meet the significance criterion in 16 of the 21 samples.

### III D: Correlations among the pretest scores

Certainly there are, as one might expect, significant positive correlations between scores on any one of the pretests and scores on the other(s) in classes in which two or more pretests were administered. In principle, this could imply the possibility that only one of the measures has a genuine association with grades while the other measures merely *appear* to have such an association due to their strong relationship to the one dominant factor. For example, the average correlation coefficients between FCI and Lawson pretest scores were in the range 0.40-0.45 and between FCI and Math they were 0.30-0.35, while those between Math and Lawson were 0.30

and 0.45 in Alg-1 and Alg-2 respectively. The standard way of testing whether one or more of the predictor variables is genuinely dominant is to carry out a multiple linear regression and see whether the coefficients of one or more of the predictor variables turn out to be not statistically significant when all of the variables are included. Indeed, as mentioned in Sec. III C, when we carried out such calculations one or more of the predictor variables often *did* turn out to be non-significant—however, there was no consistency nor any apparent pattern to these occurrences. Even within the same course at the same institution, one or the other variable (or even both) might show up as significant or non-significant from one year or one class to the next. The small size of the samples surely exacerbated this phenomenon and may well have obscured a real effect that would only be evident upon analysis of larger samples. In order to extract at least some information on this issue from our data, we turn to examination of our largest samples.

**III E: Analysis of largest samples: which pretest is most closely associated with final grade?**

In order to explore in greater depth the relative degree of association with course grades of the various pretests, we illustrate here five separate methods to determine which (if any) of the pretest scores is most closely associated with final course grade in the combined sample Alg-2 ASU-P 2022-23-24 ($N = 216$). This is the largest sample that incorporates both the Math diagnostic and the Lawson test.

1) *Correlation with grades:* The Pearson correlation coefficient between final course grade and score on the Lawson pretest was +0.28 while that for the Math pretest score was +0.35; the difference between those values is not statistically significant.
2) *Quartile odds ratio:* As illustrated in Fig. 2, the probability of obtaining a top-quartile grade for a high scorer on the Lawson pretest was 4.1 times as large as that of a low scorer on that pretest. By comparison, the probability of a top grade for a high scorer on the Math pretest was only 2.8 times as large as that of a low scorer. However, the probabilities of obtaining *bottom*-quartile grades showed the opposite pattern: low scorers on the Math pretest were 4.8 times as likely to get a bottom-quartile grade as high scorers on that pretest, while the comparable ratio for the Lawson pretest was only 2.4.
3) *Multiple linear regression:* As illustrated in Fig. S1, a multiple regression calculation yielded the equation $G = 26.363 + 0.185L + 0.295M$, where $G$, $L$, and $M$ are the percentile scores for final course grade, Lawson pretest score, and Math pretest score, respectively; both $L$ and $M$ were statistically significant predictors in this model. Since the coefficient of $M$ is substantially larger than that of $L$, the implication might be that the Math pretest score was the more “influential” of the two pretests in this case.

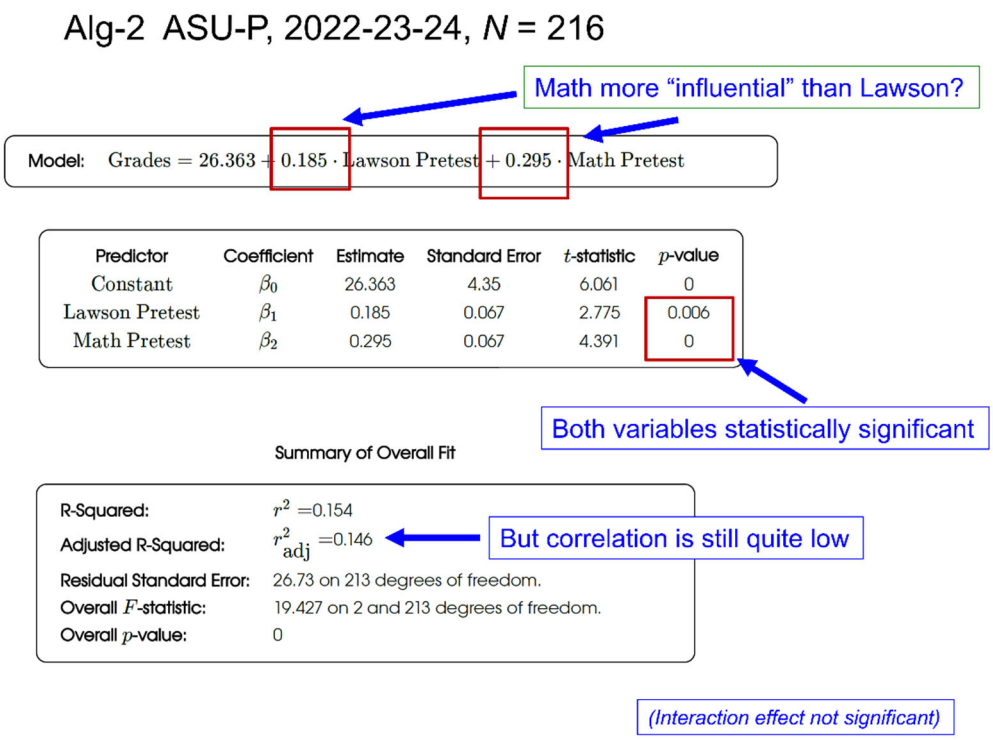



Fig. S1. Results of multiple linear regression for combined sample Alg-2 ASU-P 2022-23-24; all scores are in percentiles. Both Lawson and Math pretest scores are significant predictors of final grade points. The online calculator by Stats.Blue was used to generate the model and the graphical display.

4) *Matched pair comparison:* Four separate groups are analyzed using this method, the results of which are provided in Table SI; the four groups are (a) students in the top-quartile on Lawson pretest; (b) students in the bottom quartile on the Lawson pretest; (c) students in the top-quartile on Math pretest; (d) students in the bottom quartile on the Math pretest. In each of the four groups the same procedure was carried out, as follows: (i) Separate students with *identical* scores on the specified pretest into two separate groups of equal size, that is: those with higher scores on the other pretest, and those with lower scores on the other pretest. For example, within the Lawson top-quartile group, students scoring 84.3 on the Lawson pretest were divided into two groups of three students each, such that all students in one group had higher Math pretest scores than all students in the other group. (When an even number of students had an identical Lawson score but all different Math scores, this division was straightforward. If it was an odd number, the student with the "middle" Math score was removed from the sample. If some students with the given Lawson score also had identical Math scores, they were eliminated as necessary to form two groups of exactly equal size. If only one student had a particular Lawson score, that student was removed from the sample.) (ii) Repeat this process for all students in the (top or bottom quartile) group; the result is two separate subgroups of

identical size in which each individual student in one subgroup is matched with one in the other subgroup who has an identical score on the specified pretest (e.g., Lawson) but a different score on the *other* pretest (e.g., Math). (iii) Compare the average final grades of the higher-score subgroup to those of the lower-score subgroup. The results in Table SI show that the final grades of the higher-Math subgroups were consistently higher than those in the lower-Math subgroups (for students with identical Lawson scores), while results for the higher and lower Lawson subgroups (for students with identical Math scores) were not significantly different at the $p = 0.05$ level.

| | Top-quartile Lawson, higher Math ($N = 18$) | Top-quartile Lawson, lower Math ($N = 18$) | Significance of grade difference |
|---|---|---|---|
| **Average Lawson percentile** | 87.3 | 87.3 | |
| **Average Math percentile** | 81.6 | 31.6 | |
| **Average grade percentile** | 74.4 | 42.4 | $p < 0.001$ |

| | Bottom-quartile Lawson, higher Math ($N = 19$) | Bottom-quartile Lawson, lower Math ($N = 19$) | Significance of grade difference |
|---|---|---|---|
| **Average Lawson percentile** | 13.1 | 13.1 | |
| **Average Math percentile** | 57.1 | 16.6 | |
| **Average grade percentile** | 51.9 | 34.9 | $p < 0.02$ |

| | Top-quartile Math, higher Lawson ($N = 25$) | Top-quartile Math, lower Lawson ($N = 25$) | Significance of grade difference |
|---|---|---|---|
| **Average Math percentile** | 87.3 | 87.3 | |
| **Average Lawson percentile** | 86.5 | 39.7 | |
| **Average grade percentile** | 71.0 | 59.0 | $p < 0.09$ |

| | Bottom-quartile Math, higher Lawson ($N = 21$) | Bottom-quartile Math, lower Lawson ($N = 21$) | Significance of grade difference |
|---|---|---|---|
| **Average Math percentile** | 14.1 | 14.1 | |
| **Average Lawson percentile** | 66.6 | 17.4 | |
| **Average grade percentile** | 38.4 | 39.4 | $p > 0.9$ |

Table SI. Matched-pair comparison for combined sample Alg-2 ASU-P 2022-23-24; all scores are in percentiles. (a) Students with top-quartile Lawson scores are divided into two groups: higher Math and lower Math. Each student in the higher-Math group is matched to a student in the lower-Math group who has an identical Lawson score but a different Math score. The average grade percentiles of the two groups are compared, and the *p*-value for a *t*-test of the difference between the average grades is shown. (b) Bottom-quartile Lawson students divided into higher- and lower-Math groups, as in (a). (c) Top-quartile Math students are divided into higher- and lower-Lawson groups, matched as above. (d) Bottom-quartile Math students divided into higher- and lower-Lawson groups, matched as above.

5) *Stratify top and bottom quartiles according to scores on the "other" pretest:* Results of this calculation are shown in Fig. S2. The method is analogous to the matched pair comparison except that the groups being compared here do not necessarily have *identical* scores on one of the pretests nor are they necessarily of identical size; they only share membership in the top-quartile group. As an example, the 49 top-quartile scorers on the Lawson pretest were divided into a "higher-Math" subgroup ($N = 24$) and a "lower-Math" subgroup ($N = 25$); every person in the higher-Math subgroup had a higher Math score than anyone in the lower-Math subgroup. Students in the higher-Math subgroup were 2.3 times as likely to get a top-quartile grade as students in the lower-Math subgroup, and less likely to receive a bottom-quartile grade. (In fact, none of them received a bottom-quartile grade.) Similarly, the higher-Math subgroup among the bottom-quartile scorers on the Lawson test also had higher probability of getting a top-quartile grade and lower probability of a bottom-quartile grade than those in the lower-Math subgroup. By comparison, the results for the higher- and lower-Lawson subgroups (of the top and bottom Math quartiles) did not follow this clear pattern.

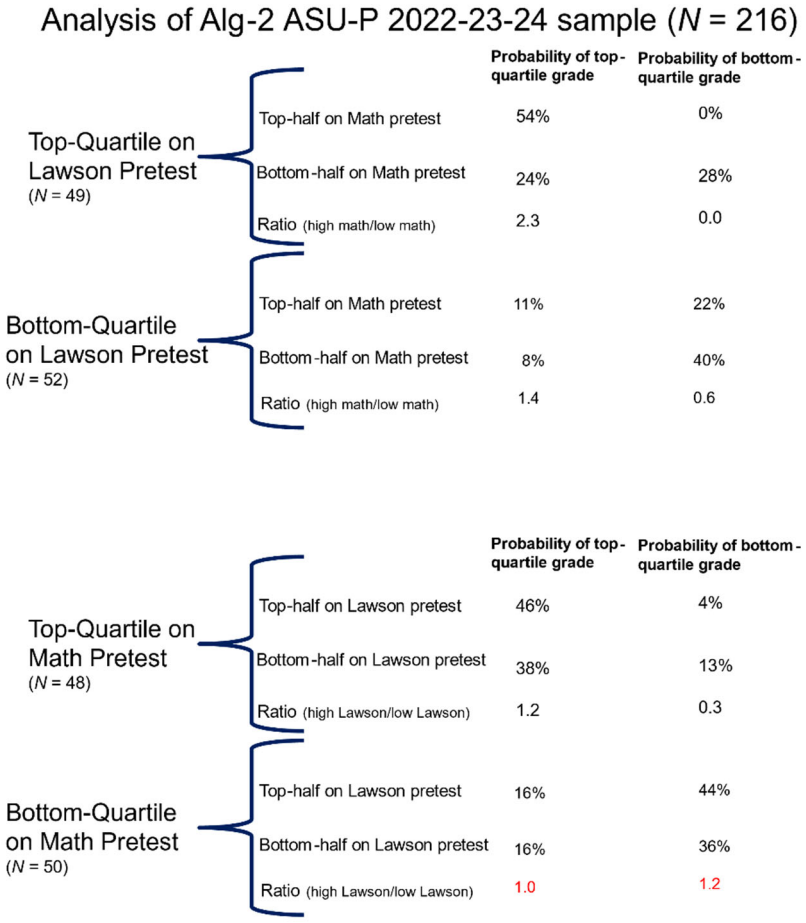



Fig. S2. Stratified sample comparison for Alg-2 ASU-P 2022-23-24. Top- and bottom-quartile grade probabilities are shown for the top-half (upper 50%) and bottom-half (lower 50%) Math scorers among both the top- and bottom-quartile Lawson groups. Also shown are the top- and bottom-quartile grade probabilities for the top-half (upper 50%) and bottom-half (lower 50%) Lawson scorers among both the top- and bottom-quartile Math groups.

Even with all these tests, the answer to the original question remains unclear. By some indications (multiple linear regression, matched pairs, and stratified samples), students' Math pretest scores appeared to be more closely associated with final grades than their Lawson pretest scores, while other indicators showed either no significant difference (correlation coefficients) or provided conflicting results (quartile odds ratios).

A similar set of calculations for the largest single sample, Alg-1 CU 2005, seemed to favor in more unambiguous fashion the Lawson pretest scores over FCI scores as the more influential of the two variables in that class. For example, the difference between the correlation coefficients (Lawson: +0.39; FCI: +0.24) was significant ($p$=0.01) and the respective coefficients in the multiple linear regression differed by factor of three (see Fig. S3). Additional analysis, shown in Fig. S4, appears to indicate that high and low Lawson scores consistently corresponded to high and low grades even when the sample was first stratified according to FCI score, while the reverse pattern (that is, FCI-grade relationship for sample stratified by Lawson score) did not hold as consistently. As the other samples were simply too small to yield meaningful results on this question, we must leave it to future research to probe it in more depth.

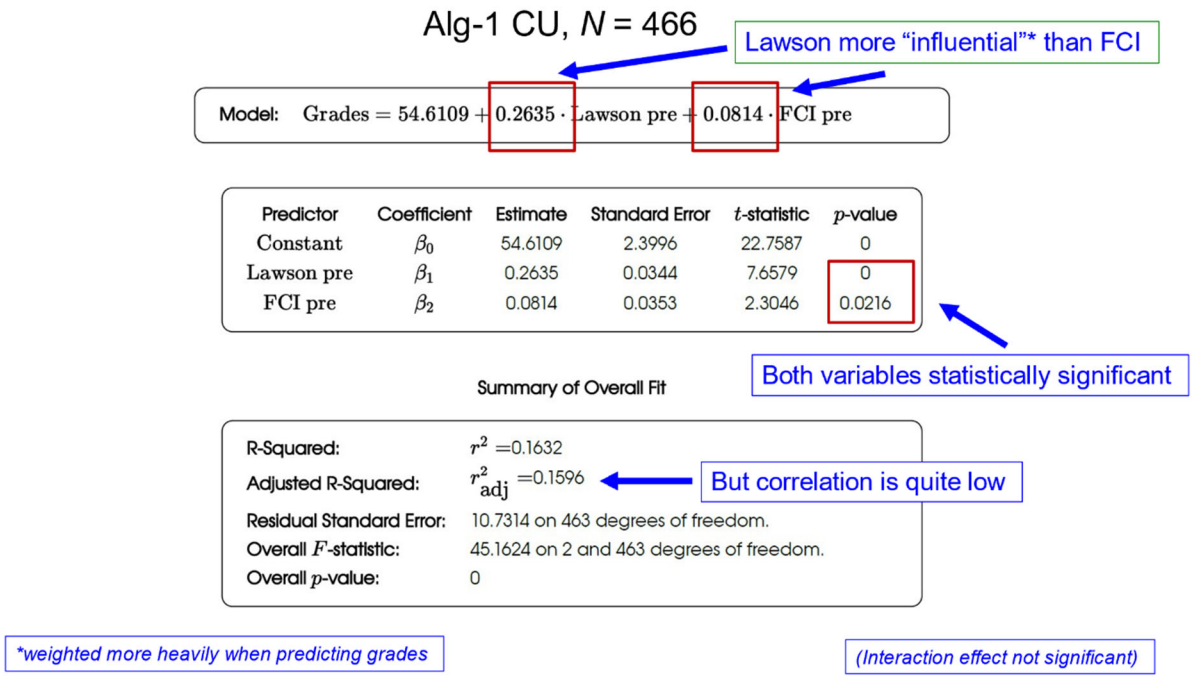



Fig. S3. Results of multiple linear regression for Alg-1 2005 CU; all scores are on a 0-100% scale. (Note that $N$ is slightly reduced because only students who took *both* pretests are included in this calculation.) Both Lawson and FCI pretest scores are significant predictors of final grade points, but Lawson appears to be more influential. The online calculator by Stats.Blue was used to generate the model and the graphical display.

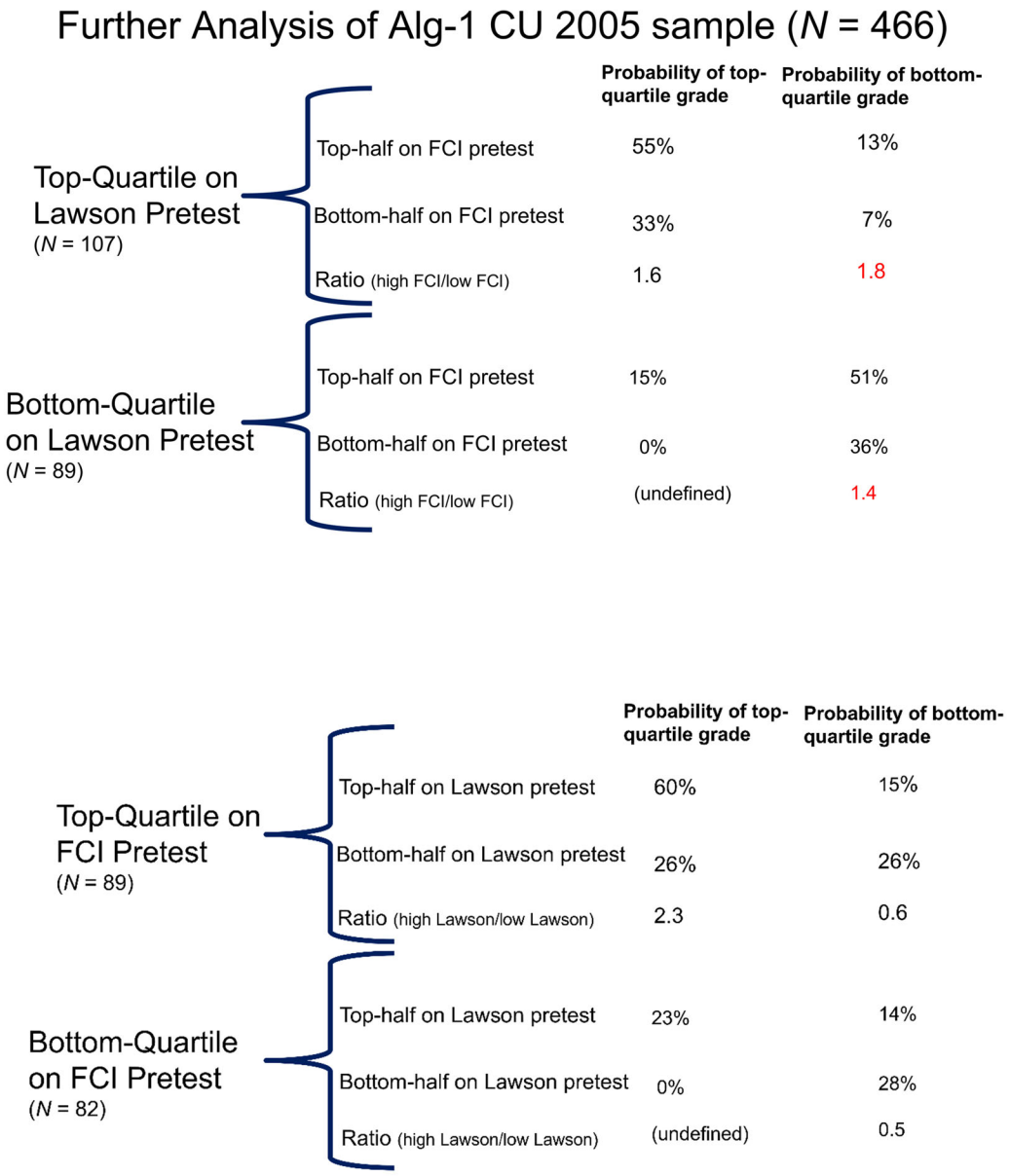



Fig. S4. Stratified sample comparison for Alg-1 2005 CU. Top- and bottom-quartile grade probabilities are shown for the top-half (upper 50%) and bottom-half (lower 50%) FCI scorers among both the top- and bottom-quartile Lawson groups. Also shown are the top- and bottom-quartile grade probabilities for the top-half (upper 50%) and bottom-half (lower 50%) Lawson scorers among both the top- and bottom-quartile FCI groups.

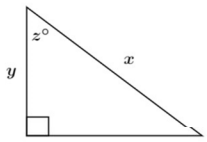
**III F: Interaction effects**

A so-called interaction effect may be present in the context of a dependent variable (grades, in the present case) that can be interpreted as responding to the influence of two or more other "independent" variables (Lawson and Math pretest scores, for example). An interaction effect would imply that the nature or strength of the functional relationship between the responding variable and one of the independent variables would vary depending on the magnitude of the *other* independent variable. For example, it might be the case that grades are more strongly dependent on (e.g., have higher correlation with) Math pretest scores for students having high Lawson scores than they are for students with low Lawson scores. A standard way to test for such an effect would be first to assume that the relationship of grade $G$ to Math score $M$ and Lawson score $L$ can be modeled as $G = a + bM + cL$ (where $a$, $b$, and $c$ are constant coefficients), but also that the Math score coefficient $b$ is *itself* linearly dependent on Lawson pretest score as in $b = d + eL$ where $d$ and $e$ are constants. (This implies that the strength of the grade-Math relationship parametrized by $b$ is itself linearly related to Lawson score.) This leads to $G = a + (d + eL)M + cL = a + dM + cL + eLM$ where $e$ is the coefficient of the "interaction" term $LM$. A coefficient $e$ that is sufficiently large indicates the presence of a significant interaction effect. However, even in situations where interaction effects may actually be present, standard statistical tests may show the interaction as "not significant" if sample sizes are too small. More broadly, the typical linear regression methods of testing for interaction effects impose rather severe assumptions on the nature of the functional relationships.

We did in fact use these standard methods to test for interaction effects in all 22 of our samples that incorporated two or more pretests, including the combined Alg-2 ASU-P 2022-23-24 sample. Only in two cases did the interaction effect test as marginally significant ($p < 0.05$) and in no case did it meet a stricter (and here more appropriate) criterion of $p < 0.01$. However, the assumptions made about the nature of the functional relationships in standard multiple linear regressions are, after all, rather restrictive, and a deeper analysis of the data suggests the possibility of an alternative interpretation. Specifically, the results presented in Table SI and Fig. S2 seem to suggest that the positive association between grades and Math scores is stronger for students with high Lawson scores than for those with low Lawson scores. In Fig. S2, the probability ratio (high Math/low Math) for obtaining high grades is higher and for low grades is lower for students with high Lawson scores, compared to those with low Lawson scores. In Table SI, the score difference (high Math vs. low Math) is larger for students with high Lawson scores than for those with low Lawson scores. Similarly, the same figures appear to show that the positive association between grades and Lawson scores may be stronger for students with high Math scores than for those with low Math scores (although the difference is not statistically significant). Together, these results suggest the possibility of an interaction effect in which the positive correlation between grades and scores on one of the pretests is larger for students who score higher on the *other* pretest. One might characterize this relationship by saying that if a student scores low on one of the pretests, their score on the *other* pretest is less predictive of their grade than it might be if they had scored high on that pretest.

# Mathematics Diagnostic Test (calculators allowed)

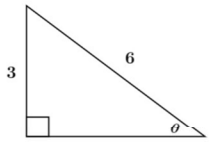
What is the length of side $x$?




A. $ycos(z°)$ D. $y/cos(z°)$ G. $cos(z°)/y$ J. $\sqrt{y^2+z^2}$
B. $ysin(z°)$ E. $y/sin(z°)$ H. $sin(z°)/y$ K. $\sqrt{z^2-y^2}$
C. $ytan(z°)$ F. $y/tan(z°)$ I. $tan(z°)/y$ L. $y/z$

(There may be more than one correct answer, but please select only ONE answer.)

What is the value of $\theta$?




A. $cos(3/6)$ D. $cos^{-1}(3/6)$ G. 30° J. 27°
B. $sin(3/6)$ E. $sin^{-1}(3/6)$ H. 45° K. 3/6
C. $tan(3/6)$ F. $tan^{-1}(3/6)$ I. 60° L. 0.524

(There may be more than one correct answer, but please select only ONE answer.)

$cos(0°) = ?$

A. 0 B. 1 C. undefined D. 0.707 E. 0.894

(There may be more than one correct answer, but please select only ONE answer.)

$sin(90°) = ?$

A. 0 B. 1 C. undefined D. 0.707 E. 0.894

(There may be more than one correct answer, but please select only ONE answer.)

$tan(0°) = ?$

A. 0 B. 1 C. undefined D. 0.707 E. 0.894

(There may be more than one correct answer, but please select only ONE answer.)

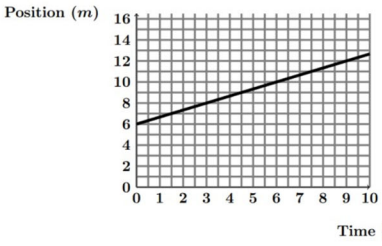
Solve for $\theta$.

$\gamma\theta + \eta = \lambda\theta + \omega$

A. $\frac{\eta+\omega}{\gamma-\lambda}$ C. $\frac{\gamma-\lambda}{\omega-\eta}$ E. $\frac{\eta-\omega}{\gamma\lambda}$ G. $\frac{\omega-\eta}{\gamma-\lambda}$ I. $\frac{\eta-\omega+\gamma}{\lambda}$
B. $\frac{\eta-\omega}{\lambda-\gamma}$ D. $\frac{\lambda-\gamma}{\eta-\omega}$ F. $\frac{\omega-\eta}{\gamma\lambda}$ H. $\frac{\omega-\eta}{\gamma+\lambda}$ J. $\frac{\omega-\eta+\lambda}{\gamma}$

(There may be more than one correct answer, but please select only ONE answer.)

What is the slope of the graph below?




A. $\frac{1}{3}$ m/s because the object moves 1 meter in 3 seconds.

B. $\frac{1}{3}$ m/s because the line rises 1 box while it goes 3 boxes in the horizontal direction.

C. $\frac{2}{3}$ m/s because the object moves 2 meters in 3 seconds.

D. $\frac{2}{3}$ m/s because the line rises 2 boxes while it goes 3 boxes in the horizontal direction.

(There may be more than one correct answer, but please select only ONE answer.)

$\left(\frac{a}{3}\right)^3 = ?$

A. $\frac{a^3}{3}$ B. $\frac{a}{27}$ C. $\frac{a^3}{27}$

(There may be more than one correct answer, but please select only ONE answer.)

$2\left(\frac{a}{b}\right) = ?$

A. $\frac{2a}{b}$ B. $\frac{2a}{2b}$ C. $\frac{a}{2b}$

(There may be more than one correct answer, but please select only ONE answer.)

$\frac{a/b}{c^2/d} = ?$

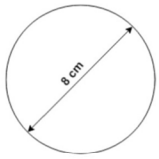
A. $\frac{ac^2}{bd}$ B. $\frac{ad}{bc^2}$ C. $\frac{bd}{ac^2}$ D. $\frac{bc^2}{ad}$

(There may be more than one correct answer, but please select only ONE answer.)

$2\left(\frac{3}{4}\right) = ?$

A. $\frac{6}{8}$ B. $\frac{12}{8}$ C. $\frac{3}{8}$ D. $\frac{3}{2}$ E. $\frac{3}{4}$

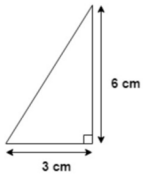
(There may be more than one correct answer, but please select only ONE answer.)




(a) Area of the circle = ?

| | | |
|---|---|---|
| A. $8\pi$ cm$^3$ | F. $8\pi$ cm$^2$ | K. $8\pi$ cm |
| B. $16\pi$ cm$^3$ | G. $16\pi$ cm$^2$ | L. $16\pi$ cm |
| C. $32\pi$ cm$^3$ | H. $32\pi$ cm$^2$ | M. $32\pi$ cm |
| D. $64\pi$ cm$^3$ | I. $64\pi$ cm$^2$ | N. $64\pi$ cm |
| E. $128\pi$ cm$^3$ | J. $128\pi$ cm$^2$ | O. $128\pi$ cm |

(There may be more than one correct answer, but please select only ONE answer.)




(b) Area of the triangle = ?

| | | |
|---|---|---|
| A. 4.5 cm$^3$ | F. 4.5 cm$^2$ | K. 4.5 cm |
| B. 9 cm$^3$ | G. 9 cm$^2$ | L. 9 cm |
| C. 12 cm$^3$ | H. 12 cm$^2$ | M. 12 cm |
| D. 18 cm$^3$ | I. 18 cm$^2$ | N. 18 cm |
| E. 36 cm$^3$ | J. 36 cm$^2$ | O. 36 cm |

(There may be more than one correct answer, but please select only ONE answer.)

Solve for x.

$\frac{3}{2} = 7x$

A. $\frac{14}{3}$ B. $\frac{3}{14}$ C. $\frac{21}{2}$ D. $\frac{21}{14}$

(There may be more than one correct answer, but please select only ONE answer.)

$v^2 = v_0^2 + 2ad$

$v_0 = 0$

$a = \frac{\Delta v}{\Delta t}$

$\Delta v = 60$

$\Delta t = 8$

$v = 30$

$d =?$

A. $d = 30$ B. $d = 60$ C. $d = 120$ D. $d = 240$ E. $d = 480$

(There may be more than one correct answer, but please select only ONE answer.)

$cy = dx$

$a - y = bx$

$x =?$

A. $\frac{ac}{d+b}$ C. $\frac{ac}{bc-d}$ E. $\frac{ac}{db}$ G. $\frac{a}{b+\frac{d}{c}}$ I. $\frac{1}{b}\left(a - \frac{d}{c}\right)$

B. $\frac{ac}{d-b}$ D. $\frac{ac}{bc+d}$ F. $\frac{a}{db}$ H. $\frac{a}{b+d}$ J. $\frac{c}{d}\left(a - b\right)$

(There may be more than one correct answer, but please select only ONE answer.)