

Student Performance on Conceptual Questions: Does Instruction Matter?

Paula R.L. Heron

*Department of Physics, Box 351560
University of Washington
Seattle WA, 98195-1560*

Abstract. As part of the tutorial component of introductory calculus-based physics at the University of Washington, students take weekly pretests that consist of conceptual questions. Pretests are so named because they precede each tutorial, but they are frequently administered *after* lecture instruction. Many variables associated with class composition and prior instruction (if any) could, in principle, affect student performance on these questions. Nonetheless, the results are often found to be “essentially the same” in all classes. With data available from a large number of classes, it is possible to characterize the typical variation quantitatively. In this paper three questions for which we have accumulated thousands of responses, from dozens of classes representing different conditions with respect to the textbook in use, the amount of prior instruction, *etc.*, serve as examples. For each question, we examine the variation in student performance across all classes. We also compare subsets categorized according to the amount of relevant prior instruction each class had received. A preliminary analysis suggests that the variation in performance is essentially random. No statistically significant difference is observed between results obtained before relevant instruction begins and after it has been completed. The results provide evidence that exposure to concepts in lecture and textbook is not sufficient to ensure an improvement in performance on questions that require qualitative reasoning.

Keywords: physics education research, conceptual understanding

PACS: 01.40.Fk, 01.40.gb

INTRODUCTION

For more than 20 years, weekly “pretests” have been part of introductory calculus-based physics at the University of Washington (UW). Pretests are so named because they precede each tutorial, but they are frequently administered *after* lecture instruction. The pretests play a key role in the implementation of *Tutorials in Introductory Physics*, a set of instructional materials the Physics Education Group at the UW has developed to supplement instruction by lecture and textbook [1]. Up to 1400 students take the introductory calculus-based course each quarter and some pretests have been given many times over the past two decades. Thus the responses constitute a large data set.

Many articles by our group have asserted that on the types of conceptual questions that comprise the pretests, student performance is “essentially the same” before and after instruction in lecture (but not tutorial). There are two claims implicit in this statement: (1) there is no systematic variation due to instruction and (2) the range of observed results is, in some sense, small. This paper examines these claims in greater detail through a retrospective analysis of results from three sample questions. We look at the overall variation between sections, and then compare results obtained at different stages of instruction.

ABOUT THE DATA SOURCE

Tutorials are part of all three parts of the introductory calculus-based sequence at UW: Mechanics, E&M, and Waves & Optics. Each is offered every academic quarter, often in more than one lecture section. Each section (also referred to here as a class) enrolls up to 225 students and is taught by a different faculty member in the Physics Department.

Since 2000, the pretests have been administered online. Students are asked to select answers to several conceptual questions from a menu of choices and to type brief explanations in a text box. They are given credit, whether or not their answers are correct. (Periodic spot-checking helps ensure that students take the pretests seriously.) Students have 15 minutes to take the pretest, which they can do at any time during a roughly 48-hour period that starts after their Friday lecture and ends before their Monday lecture.

Students in the Honors section of the course also take pretests, but we exclude those data. Other than that course, we have no reason to believe that that enrollment in any given quarter or section is biased in favor of any particular major.

Although we refer to them as “pretests” it is important to emphasize that in many cases they follow instruction on the relevant concepts in lecture.

However, because of periodic changes to the syllabus, the amount of prior instruction on the topic varies widely: sometimes there has been none; other times all of the relevant lectures, homework problems, and laboratory experiments have been finished. The nature of the instruction also differs because the course instructors – while following a common syllabus – prepare their own lectures and choose their own sample problems, derivations, and lecture demonstrations. Recently most have also used “clicker” questions of their own design. In addition, the textbook from which readings and homework problems are assigned changes every few years.

In this paper, three conceptual questions serve as examples. (See Figure 1.) They were chosen because, as shown below, the variation in performance from class to class appears to follow different patterns.

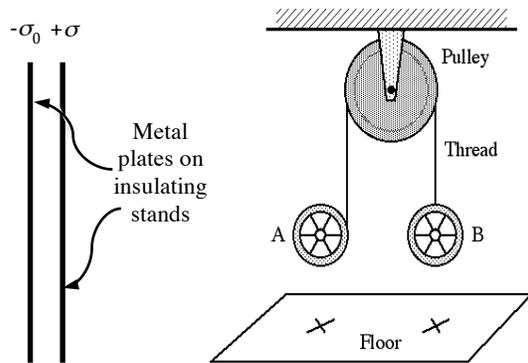


FIGURE 1. Figures shown with questions CAP1, CAP2 (left) and DRB1 (right).

Two questions are associated with a tutorial on capacitance. In CAP1 students are asked whether the

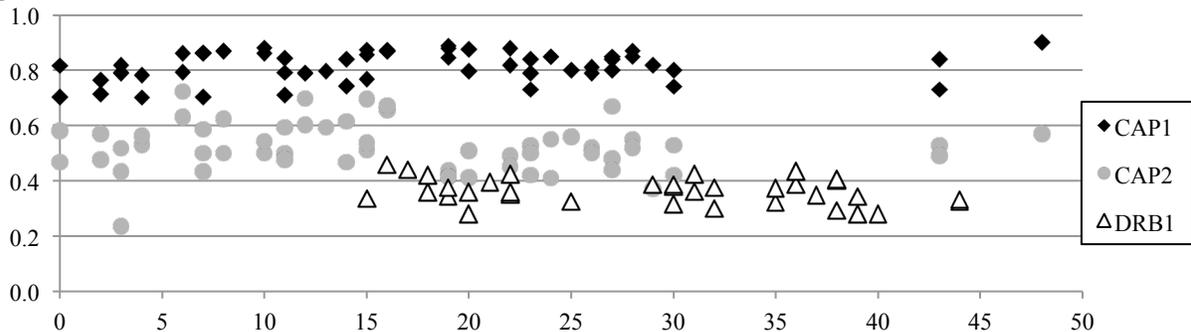


FIGURE 2. Results for all three questions. A total of 57 lecture sections (5171 students) responded to CAP1 and CAP2; 35 sections (4012 students) responded to DRB1. The horizontal axis represents the quarter in which the question was asked. Often a given question was asked in more than one course section during that quarter; each class is shown separately.

The simplest model is that a proportion p of this population of students has probability 1 of giving the correct answer to a given question and a proportion $(1 - p)$ has probability 0. If each class is of the same

charge density on each plate increases, decreases, or remains the same as the plates are moved closer together. In CAP2 they are asked the same question about the capacitance of the pair of plates. Question DRB1 is associated with a tutorial on the dynamics of rigid bodies. Students are asked which spool will hit the ground first if they are released from rest at the same instant. [2]

RESULTS

The results for all three questions (arranged chronologically) are shown in Figure 2. The vertical axis represents performance π_i of class i on a given question:

$$\pi_i = \frac{n_i^{corr}}{n_i^{total}} \quad (1)$$

where n_i^{corr} is the number of correct answers and n_i^{total} is the total number of answers received. As shown, performance on CAP1 (diamonds) is generally higher than either CAP2 (circles) or DRB1 (triangles).

Nature Of The Distributions

If none of the variables associated with instruction (e.g., instructor, textbook, extent of coverage of the relevant topic) has an effect on student ability to answer a given question correctly (and there is no systematic variation in class composition), then each class represents a random sample (of roughly the same size) from the population of UW students intending to major in fields that require calculus-based physics (with the exclusion mentioned above). Therefore π should be a random variable.

size, the set of values $\{n_i^{corr}\}$ should follow a binomial distribution. This model can be tested by generating predicted distributions for each question, assuming a class size of 100 (the average in our data set). The

proportion p is estimated to be the mean of the set of observed values $\{n_i^{corr}\}$:

$$\bar{\pi} = \frac{\sum_i \pi_i}{N} \quad (2)$$

where N is the number of classes in which the question was asked. These distributions are shown in Figure 3 along with the actual distributions. (For simplicity the size of each actual class is taken to be 100 so that $n_i^{corr} = 100 \pi_i$.) [3] Visual inspection suggests that the results for DRB1 are a good match. (The means are the same by design; the shapes of the distribution are significant.)

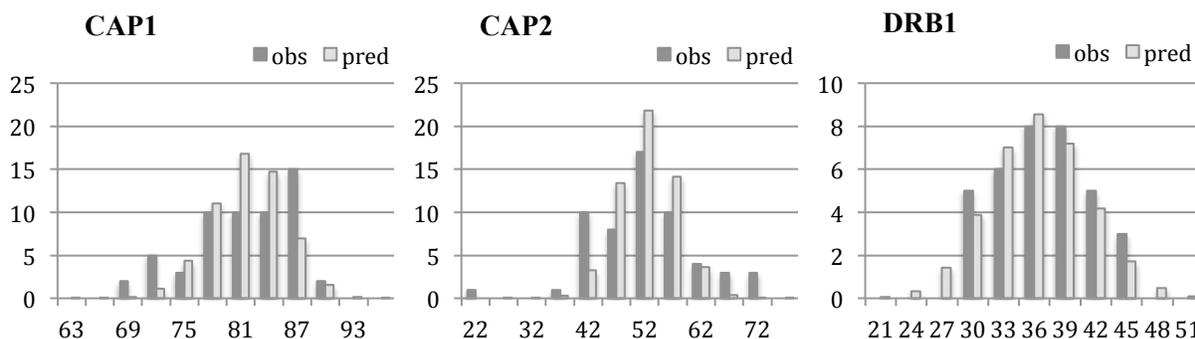


FIGURE 3. Histograms with actual distribution (obs) and predicted binomial distributions (pred).

TABLE 1. Actual and predicted standard deviations. Predictions are based on the normal approximation to the binomial distribution and assume 100 students in each class.

	CAP1	CAP2	DRB1
actual	5.3	8.9	4.7
predicted	3.9	5.0	4.8

Even though the results of CAP1 and CAP2 do not support the simple model described above, they still appear to be randomly distributed. Using a “normal probability plot correlation coefficient test” and choosing $\alpha = 0.05$ the null hypothesis (that the distributions are normal) is not rejected for DRB1 or CAP2, while choosing $\alpha = 0.01$ it is not rejected for CAP1 (as shown, the distribution for CAP1 is somewhat skewed, as expected given the high mean).

Effect Of Instruction

Figures 2 and 3 include all classes, regardless of the timing of the questions with respect to discussion of the relevant concepts in lecture. For CAP1 and CAP2, all classes had completed instruction on Coulomb’s law, the electric field due to continuous charge distributions (*e.g.*, infinite lines and sheets), Gauss’s law, and potential, but were at different stages of instruction on capacitance. While CAP1 can be

To put the comparison on a more quantitative basis, the binomial distribution can be approximated by a normal distribution (because N is large) and the predicted and actual standard deviations compared. The predicted standard deviation for the normal approximation to the binomial distribution is $\sqrt{np(1-p)}$. The sample standard deviation of the set of observed values $\{n_i^{corr}\}$ can be calculated as:

$$\sigma = \sqrt{\frac{\sum_i (n_i^{corr} - \bar{n}^{corr})^2}{N-1}} \quad (2)$$

As shown in Table I, the predicted and actual values are similar for DRB1, but not for CAP1 and CAP2.

answered without any knowledge of capacitance (so in that sense all results were obtained *after* instruction), a correct response for CAP2 requires at least the definition of capacitance. DRB1 was given at various stages of instruction on rigid-body dynamics; however Newton’s second law is sufficient.

Figure 4 shows the results for classes in which relevant instruction: (0) had not yet begun, (1) was underway, or (2) had been completed. As suggested by the results above (which indicate random variation) there is significant overlap.

Table 2 gives the corresponding averages. Both t -tests and (non-parametric) permutation tests indicate that only two differences are statistically significant ($\alpha = 0.05$): for CAP1, the “during instruction” average is slightly *lower* than the “before instruction” average; for CAP2, it is slightly *higher*. In no case is there a statistically significant difference between the “before instruction” and “after instruction” results.

TABLE 2. Average performance (with standard deviation) at different stages of instruction.

	CAP1		CAP2		DRB1	
	mean	sd	mean	sd	mean	sd
before	0.85	0.03	0.47	0.07	0.39	0.03
during	0.81	0.06	0.54	0.09	0.36	0.05
after	0.83	0.05	0.50	0.06	0.36	0.05
ALL	0.82	0.05	0.52	0.09	0.36	0.05

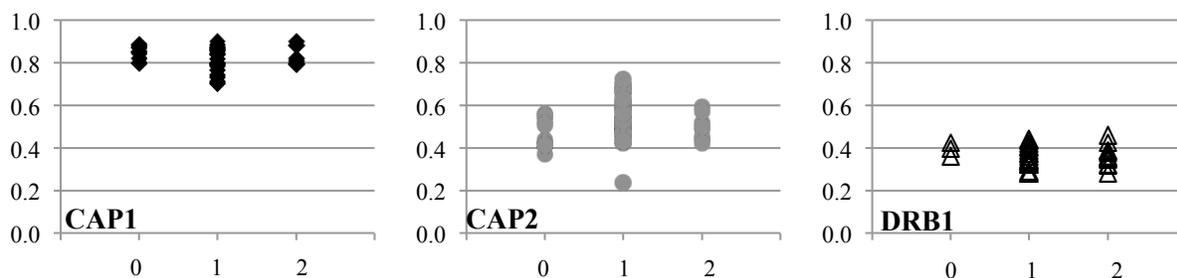


FIGURE 4. Results for classes in which instruction: (0) had not begun, (1) was underway, (2) had been completed.

DISCUSSION

The standard deviations shown in Table 2 are an indication of the range over which performance varies from class to class. For CAPI and DRB2, the bulk of the results fall within 5% of the mean; the range for CAP2 is greater. The former are typical, and help explain a practice that has been common in our group: rounding to the nearest 5% when quoting averages.

The variation in performance from class to class for the three questions differs in the degree to which it can be explained by a simple model. The results of one question (DRB1) can be explained by assuming each class is a random sample from a population consisting of two types of students. A more elaborate (and realistic) model in which the probability of answering correctly is a continuous variable (as in item response theory) might explain the results of the other two.

The very slight variations in average performance before and during instruction, if robust, may also help explain the results of CAPI and CAP2. While the overall distributions appear normal or nearly so, it may be the case that there are overlapping distributions, with means so similar that the bi- (or multi-) modality is obscured. Sayre *et al.* recently reported temporary instruction-related changes in performance on questions about Newton's third law.[4] Their results support the conclusion that the differences reported here are real (albeit so small as to have debatable significance in terms of instruction).

The apparent variation of performance with the amount of prior instruction might also be random. In other words, if we examine enough questions we might expect to see some with slight increases in performance at some stages, some with slight decreases. However, if these differences prove to be robust, they may reflect the different types of reasoning demanded by the three questions. CAPI requires simple commitment to the conservation of charge. CAP2 requires more formal reasoning with a multivariable relationship. DRB1 can be answered on the basis of experience, or by using Newton's laws.

CONCLUSION

The results reported here provide evidence that exposure to relevant concepts in lecture and textbook is not sufficient to ensure an improvement in performance on conceptual questions. The results also suggest that other variables associated with instruction (instructor, textbook, *etc.*) either have competing effects, or no effects, such that performance varies essentially randomly. More detailed analysis, which is forthcoming, could determine which explanation is appropriate. In any case, the (near) normality of the overall distributions supports the mean and standard deviation as good statistics, especially when applied in assessing the impact of an instructional intervention. The narrow range over which performance varies has implications for obtaining good estimates of population means and standard deviations when the number of classes is low. This issue is the subject of future work.

ACKNOWLEDGMENTS

The author appreciates the contributions of other members of the Physics Education Group and the support of the NSF through grant DUE 0088840.

REFERENCES

1. L.C. McDermott, P.S. Shaffer and the PEG at UW, *Tutorials in Introductory Physics*, Upper Saddle River: Pearson, 2002.
2. H.G. Close, L.S. Gomez, and P.R.L. Heron, "Student understanding of the application of Newton's second law to rotating rigid bodies," to appear in *Am. J. Phys.* (2012).
3. The use of a standardized class size of 100 may make a slight difference in calculations statistics reported here but does not alter the basic results.
4. E.C. Sayre, S.V. Franklin, S. Dymek, J. Clark, and Y. Sun, "Learning, retention, and forgetting of Newton's third law throughout university physics," *Phys. Rev. ST Physics Ed. Research* **8**, 010116 (2012).