Issues Related to Data Analysis and Quantitative Methods in PER

David E. Meltzer
Department of Physics & Astronomy, Iowa State University, Ames, IA 50011

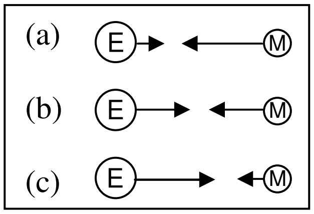
A variety of issues are always relevant (either explicitly or implicitly) in analysis of quantitative data in Physics Education Research. Some specific examples are discussed.

There are a number of issues that always arise, implicitly or explicitly, when conducting quantitative research and carrying out data analysis in Physics Education Research. (Most are relevant for qualitative research as well.)

I. Validity. Broadly speaking, validity refers to the degree to which the conclusions of an investigation truthfully and accurately respond to some specific research questions. Among the particular issues that may arise is: Does your instrument provide data that could actually answer your research question? A common flaw is that the instrument (or test item) is not sufficiently focused, in this sense: To try to answer the question, "Do students understand concept *A*?" the test item (or test instrument) requires knowledge of concepts *A*, *B*, and *C*. Here, *B* and/or *C* might correspond to specific mathematical tools or formal representations. A related question that might arise is: Is your interpretation of the data an accurate representation of students' knowledge?

For example, consider how one might assess students' knowledge of Newton's third law in the context of gravitational forces. At Iowa State I have given a quiz on gravitation on the second day of class for five consecutive years. (The course is the second semester of the algebra-based general physics sequence, focusing on electricity and magnetism. All students in this course have completed their study of mechanics.) Question #1 on the quiz asks whether the magnitude of the gravitational force exerted by the sun on the earth is larger than, the same as, or smaller than the magnitude of force exerted by the earth on the sun. (This question uses words, but no diagrams or equations.) The correct answer ("the same") was given by 10-23% of the students (representing the low and high scores among the five classes). The most popular response by far was "larger," and it was given by 70-83% of all students.

On the very same quiz, Question #8 asks the students to choose a vector diagram that most closely represents the gravitational forces that the earth and moon exert on each other. The three most popular choices are shown in the figure below.



The correct answer "*b*" was given by 6-12% of students. In each of the five independent administrations of the quiz, the proportion of correct responses on Question #8 was about half that on Question #1 (0.43, 0.60, 0.59, 0.50, and 0.50). The implication seems to be that Question #8 was measuring not only students' knowledge of Newton's third law of motion and law of gravitation, but also (in part) students' understanding of vector diagrams. This conclusion is considerably strengthened by the fact that 34-47% of students gave answer "*c*" on Question #8 [answer "*a*": 43-55%]. The "*c*" response corresponds to the force exerted by the more massive object having the *smaller* magnitude, a response that was given by only 3-6% of the same students on Question #1. We see, then, that the validity of two inferences that *might* have been drawn from the results on Question #8 are thrown into question: (1) the proportion of students who misunderstood Newton's third law, and (2) the proportion who believed that in a gravitational interaction involving two masses, the more massive object exerts the *smaller* magnitude force. Although a more definitive analysis of students' reasoning on these questions must await examination of interview data (currently underway), it seems clear that the validity of conclusions that might have been based on only one of these test items would be very uncertain.

The lesson to be drawn from this example is simply the ever-present need to be cautious in collecting and interpreting PER data. Although writers of diagnostic instruments and test items must *always* make some assumptions regarding the previous knowledge of the students being tested, it is important to (1) be aware of what specific assumptions are being made, and (2) have some sound basis (e.g., previous investigation) for believing that the assumptions are accurate.

Another threat to validity of interpretations of test data is associated with analysis of students' answers without regard for explanations of their reasoning. Although there are many good practical reasons for employing diagnostic instruments that yield "answer only" data without students' explanations, it is important for researchers to be aware of possible pitfalls in the data analysis. These dangers are associated most particularly with attempts to draw conclusions from only one or a small number of test items. For example, in a study at the University of Washington [1], students were asked to compare the changes in kinetic energy and momentum of two objects of different mass, acted upon by the same force. For both of these comparisons, the proportion of correct responses observed when ignoring students' explanations was substantially higher than when answers were judged correct only when accompanied by a correct explanation. (KE comparison: 45-65% correct vs. 30-35% correct; momentum comparison: 55-80% correct vs. 45-50% correct.) Many other researchers have reported anecdotal evidence that supports the conclusion suggested by this study, that is, that data regarding students' explanations of their reasoning (whether in written or verbal form) very substantially strengthen the potential validity of conclusions drawn from any given investigation.

II. Reliability. Reliability refers to the consistency of results produced by a specific instrument or investigative protocol. It is related to validity in the sense that an *unreliable* instrument is very unlikely to lead to valid conclusions about a research question. Reliability encompasses several distinct concepts: (1) Is the instrument internally consistent, that is, do different components of the instrument measure (more or less) the same property? This may be investigated with such measures as KR-20 or Cronbach's alpha [2]. Note that an instrument might well be designed that *intentionally* measures two or more distinct conceptual areas and therefore might not be expected to yield similar results on different subsections. (2) If you made the same measurement again (with all conditions apparently identical), would your instruments yield the same result? If a particular test item or a small set of items deal with a concept of which students have little or no knowledge, responses tend to be random. Therefore, even two consecutive administrations of the same instrument might yield substantially different results and analysis should take that into consideration. (3) Would minor variations in your test items (e.g., slight contextual or representational changes, or alterations in question format) lead to large variations in results?

For example: Schecker and Gerdes [3] reported significant differences in student responses to certain FCI questions when the questions were posed in slightly different physical contexts, i.e., a soccer ball instead of a golf ball, or a vertical pistol shot instead of a steel ball thrown upward. Steinberg and Sabella [4] administered final-exam problems in free-response format that were similar to several FCI questions. They found that in some cases, there were significant differences in percent correct responses between the final-exam questions and corresponding FCI items (administered post-instruction) for students who took both tests. In the example discussed in Section I above, two very similar questions on gravitation posed in different representational forms yielded significantly different results, suggesting that the reliability of an instrument that depended on only one or the other type of question might be compromised. With regard to multiple-choice exams, Rebello and Zollman [5] have provided evidence that even well-validated multiple-choice questions might miss categories of responses that students would offer were the questions posed in free-response format. They also show that in some cases, the specific selection of distracters provided to students can significantly affect the proportion of correct responses.

Again, it should be emphasized that researchers are always forced to make some assumptions regarding the reliability of their instruments and methods. Nonetheless, some efforts – however informal – should be made to gauge the reliability of any particular investigative protocol.

More generally, diagnostic items that omit students' explanations may have their reliability threatened for that reason alone. In the University of Washington study discussed above [1], both questions (i.e., KE comparison and momentum comparison) were posed in two separate variants: one in which the different objects experienced forces for the same time period, and one in which the time periods differed. Remarkably, the proportion of correct responses when explanations were required was nearly identical for the two variants (KE: 35% and 30%; momentum: 50% and 45%). However, when explanations were ignored, results on the two variants were significantly different (KE: 65% and 45%; momentum: 80% and 55%). This suggests that reliability, and not merely validity, may be strongly dependent on consideration of student explanation data.

III. Statistical Significance. Before drawing any conclusions from one's data it might be helpful to ask whether there is a substantial probability (10% or more) that your result might have occurred purely by chance. Do you have a measure of variance, or can one be estimated? If standard deviations are available a *t*-test (or similar measures) could be used to assess significance of differences in sample means. If not, an assumption of binomial distribution might be made and a test for difference between binomial proportions could be applied [6].

If many individual variables or inter-sample differences are being tested for significance, then substantial deviations from "null hypothesis" values may be expected to occur, purely by chance, for some tested items. For instance, if 100 different sample means are compared, random fluctuations would dictate that several are likely to show a two-sigma ($p = 0.05$) effect (i.e., means separated by two or more standard errors).

Another important consideration is that the sample size being utilized may be inadequate to yield a statistically significant result for the specific effect being investigated. In that case, failure to observe a difference between control and experimental groups may *not* imply non-existence of a treatment effect, but merely that the sample size used or the experimental protocol employed is inadequate to demonstrate the existence of the effect at an acceptable level of statistical significance.

IV. Pedagogical Significance. Is the observed effect likely to be of practical significance in the classroom? Are there cost-benefit relationships implied in the magnitude of the effect [7]? Even if an effect is statistically significant (e.g., large "effect size" [8]) the actual learning gains (as measured for instance by Hake's *g* [8, 9]) might be small and of limited practical pedagogical interest.

V. Representativeness of Sample. Is your student sample representative of the larger group from which it is (implicitly or explicitly) drawn? Are samples from the different student groups that are being compared equivalent in all respects except for the variable being investigated? If sample selection is truly random the expectation is that the answer to both of these questions should be "yes." In random samples that are sufficiently large, the probability that both answers actually *are* "yes" is very high. However, samples are rarely "sufficiently large" nor, for that matter, truly randomly selected. In that case one must consider which relevant population variables may differ among the various student samples, for example: demographic makeup, previous preparation, pre-instruction knowledge, etc. Although some measures of learning gain such as Hake's *g* explicitly incorporate normalization to reduce the dependence on pretest scores [8, 9], so-called "hidden variables" such as mathematics preparation, gender, spatial visualization ability, reasoning ability, etc. may nonetheless exert an influence for which account should be taken [8, 10]. Even more subtle variables such as whether students are enrolled in an "on-sequence" or "off-sequence" course might have an effect [11].

One should always ask: How have you controlled variables that might be relevant? Have you done random selection? If not, what alternatives were used? In any case, what is the basis for believing that the different population samples being compared are equivalent except for the treatment being tested?

VI. Reproducibility. Just because you saw an effect in one PER experiment does not necessarily mean you will observe it again. In physics, all groups of electrons in identical states are completely equivalent. In PER, different groups of students are never in identical states and are never truly completely equivalent. This reality requires answers to questions such as these: Did you repeat the experiment? Did anybody else repeat the experiment? Are your results substantially different from what others have observed, or are they otherwise very surprising? If so, better check again!

It is important to keep in mind that PER necessarily deals with many variables that are often difficult (and sometimes impossible) either to identify or to control (or both), e.g.: student demographics, instructor style, course logistics, issues of validity and reliability of diagnostic instruments, etc. Moreover, students' mental models of physics concepts are often complex and incorporate overlapping and frequently conflicting themes. Therefore, students' responses to different (though related) questions may be highly variable. Largely due to this assortment of variables, fluctuations from one PER data run to the next tend to be large (and, of course, each data run may require an entire academic quarter or semester). This inherently large scale of fluctuations substantially increases the importance of replication in PER investigations in comparison, for instance, to more traditional physics research. Even investigations that yield large treatment effects with high statistical significance should probably be replicated by the original research group at the same institution, and/or by other researchers working at different institutions with diverse student populations.

SUMMARY

Although the issues that are discussed here often get no explicit attention in Physics Education Research papers and presentations, I believe that PER investigators should formulate responses – at least implicitly and approximately – to all questions of this type. Substantial neglect of one or more of these issues can threaten the validity and usefulness of the results of an investigation, and vitiate the product of hundreds of hours of laborious study.

I am grateful for discussions with Leith Allen and R. Hake. This material is based upon work supported by the National Science Foundation under Grant Number DUE-9981140 and Grant Number REC-0206683.

REFERENCES

1. T. O'Brien Pride, S. Vokos, and L. C. McDermott, "The challenge of matching learning assessments to teaching goals: An example from the work-energy and impulse-momentum theorems." Am. J. Phys. **66**, 147 (1998).

2. W. R. Borg and M. D. Gall, *Educational Research*, *An Introduction* (Longman, New York, 1989), 5th ed., pp. 260-261.

3. H. Schecker and J. Gerdes, "Messung von Konzeptualisierungsfähigkeit in der Mechanik: Zur Aussagekraft des FCI," Zeitschrift für Didaktik der Naturwissenschaften **5** (1), 75-89 (1999).

4. R. Steinberg and M. Sabella, "Performance on multiple-choice diagnostics and complementary exam problems," Phys. Teach. **35**, 150 (1997).

5. N. S. Rebello and D. A. Zollman, "The effect of distracters on student performance on the Force Concept Inventory," Am. J. Phys. (Phys. Educ. Res. section) *in press*.

6. J. L. Devore, *Probability and Statistics for Engineering and the Sciences*, 2nd ed. (Brooks/Cole, Monterrey, CA, 1987), Ch. 9.2, 9.4.

7. R. R. Hake, "Effect sizes and economic cost-benefit," AERA-D post at <http://lists.asu.edu/cgi-bin/wa?A2=ind0207&L=aera-d&F=&S=&P=3921>.

8. R. R. Hake, "Relationship of individual student normalized learning gains in mechanics with gender, high-school physics, and pretest scores on mathematics and spatial visualization," <http://www.physics.indiana.edu/~hake/index.html>.

9. R. R. Hake, "Interactive engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," Am. J. Phys. **66,** 64-74 (1998); <http://www.physics.indiana.edu/~sdi/>.

10. D. 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) and at:
<http://www.physics.iastate.edu/per/index.html>.

11. N-L Nguyen and D. E. Meltzer, "Initial understanding of vector concepts among students in introductory physics courses," Am. J. Phys. (Phys. Educ. Res. section), accepted for publication, <http://www.physics.iastate.edu/per/index.html>.