Research Methodologies in Science Education

Strategies for Productive Assessment

Julie C. Libarkin

Science Education Department, Harvard-Smithsonian Center for Astrophysics, 60 Garden St., MS-71, Cambridge, MA 02138, jlibarki@cfa.harvard.edu

Josepha P. Kurdziel

Teaching and Teacher Education, Education Room 703, University of Arizona, Tucson, AZ 85721, kurdziel@u.arizona.edu

Why would an instructor, department, or institution want to "assess" their courses? What does it mean to assess a course? How would they go about performing this assessment? These are questions that we are routinely asked at professional meetings by scientists interested in understanding how to improve their teaching. Here we lay out a few guidelines that will help you interpret published education research, apply this research to your own classroom, or engage in your own research endeavors. Because the techniques used in science education research have more in common with the behavioral sciences than the physical sciences, it is often difficult for those in the "hard" sciences to interpret the education literature. This column will serve as a medium for highlighting the most important, useful, or easily applied techniques, with additional guidelines for what to look for in your own literature reviews.

Basic Principles of Science Education Research. Before you can get down to the business of assessing a course, curriculum, or teaching method, it is important to carefully lay out the framework in which the assessment will be carried out (Terenzini, 1989). We believe the following questions, if answered carefully and honestly, will go a long way to ensuring that your results are useful, not only for yourself, but for the community at large (Terenzini, 1989; Johnson, 1997; Shea, 1999). We have developed an imaginary case study to help illustrate how assessment can be accomplished. Professor Armstrong's efforts mirror similar research endeavors we have conducted ourselves.

What Are You Trying to Find Out? Are you interested in determining the effect a course has on student learning? If so, you must pinpoint the exact facet of learning you expect to be affected by your course. Your research question, just as in the sciences, must be focused and specific. There are a number of possible student outcomes: content knowledge acquisition, skills development, changes in attitudes/values/beliefs, and long-term behavioral outcomes (Ewell, 1987). Which of these do you believe will change for your students as a result of participation in your course?

Professor Armstrong teaches an introductory geology course and has begun to use undergraduate peer teachers in his classroom. He wants to find out if this addition to his course has helped his students achieve any of his course goals. However, he isn't sure which goals will be affected by the peer teachers. He is fairly confident that communication skills will improve, but this is a secondary goal. His course syllabus states that the course is designed to improve "student attitudes towards science, content knowledge, ability to evaluate scientific issues in the news, and ability to apply geological principles to understanding the Earth". He decides that student attitudes might change in response to peer teaching, as he has heard from other professors that students seem more positive about science when they get to discuss it with other students.

Has This Type of Study Been Conducted Before? A literature search can go a long way towards clarifying your research objectives. Several useful education reference databases exist, including ERIC, EbscoHost, and thegate- way.org. Additionally, most discipline specific databases contain education references, including INSPEC (physics), Biosis (biology), and GeoRef (geology). Science education, like many disciplines, is cyclic, so many of the questions you are asking today may have been asked in the past. It is always helpful to be aware of existing research, to ensure that your research question will be both useful and interesting.

Professor Armstrong isn't sure if anyone has ever looked at student attitudes before, and he isn't really sure how to find out. He searches GeoRef and finds several articles mentioning the word 'attitude'. He calls the local teaching center and they send him a couple more references. Armed with this literature, he feels ready to proceed with his assessment.

Professor Armstrong doesn't know it, but he has missed out on most of the research in attitudes conducted in other disciplines. A search using ERIC or another education database would have uncovered many references Dr. Armstrong missed in his initial literature search.

How Will the Assessment be Done? Multiple methods must be used to ensure that you are gathering data that will ultimately be useful in answering a specific research question. Many researchers recognize that they have some assessment measures already in hand. For instance, the literature contains many examples of analyses that rely upon student evaluations or grade distributions (i.e., Bair, 2000; Muehlberger and Boyer, 1961). However, because the effects of your course may not be apparent in these readily available sources, other methods should be used. Although any source of information can provide poten-

tially useful information, it is important to coordinate your research plan with your overall objectives. Finally, although this can be tricky, the validity and reliability of assessment instruments should be established before they are used.

Professor Armstrong knows that at the end of each semester his students will complete an evaluation of his course. He decides to use these evaluations as one way to test his course's effect on student attitudes. On top of this, one of the articles Professor Armstrong found contains a five question survey about student attitudes towards science. The questions are written so that students are asked to indicate their level of agreement or disagreement with a statement. This seems like a good test to use, because answers can be scored, and the test average can be used to determine if students have positive or negative attitudes. The survey Professor Armstrong decides to use is called a Likert-scale and looks like this:

> Indicate whether you strongly agree, agree, are neutral, disagree or strongly disagree with each of the following statements.
>
> 1) I like to learn by experimenting rather than just being told about things.
> 2) I like to read articles about science.
> 3) I think science classes are more interesting than other types of classes.
> 4) Science classes are boring.
> 5) I think science is interesting.

The survey Professor Armstrong has decided to use may or may not be valid, and will be examined more closely below.

When Will This Assessment Take Place and Who Will You Be Assessing? Conceptual change takes time and it is possible that this change will not be recognizable immediately after completion of a course (Terenzini, 1989). You may find it useful to collect data before your course begins, after your course is completed, and then again several semesters or even years later. It is very important to consider the scale over which you believe change will occur so that all the necessary data are collected. Additionally, it may be important to use a control group, such as another course that differs from yours in a significant way. Finally, it is important to account for any variability within your students. Are there subgroups within your course that may respond differently to instruction? The literature search may reveal some possible differences in student response, and it is important to gather all of the necessary information while the students are on hand. Unlike a geologic sampling locality, it is not always easy to revisit your students.

Professor Armstrong decides to give this survey to his students at the beginning and end of the semester. Comparing the course average from both administrations of the survey should help him determine if his course has any effect on student attitudes. He also convinces a colleague, Professor Hyatt, to test her students. Professor Hyatt does not use peer teachers in her classroom, so they hope to be able to use this survey to compare the two different course structures (Figure 1).

Who Will Be Analyzing the Data? Finally, do you plan to conduct the data analysis yourself or ask someone to perform this service for you? You may find that education faculty or graduate students on your campus are eager to collaborate, by either teaching you the necessary statistical techniques or working with you. There are also a number of published resources that do an excellent job of explaining the basics (i.e., Nitko, 1996).

Professors Hyatt and Armstrong both have many years of experience teaching in introductory classrooms. Based on this experience, they feel confident in their ability to score and interpret the attitude surveys.

ASSESSING YOUR COURSE.

The semester is now over and the two professors sit down to analyze and interpret their data. They each had ~80 students in their course, although several students didn't take the pre-test, post-test, or both. Professor Armstrong has 70 students who took both tests and Professor Hyatt has 68. They have scored the test like this:

strongly agree=1, agree=0.75, neutral=0.50, disagree=0.25, strongly disagree=0

They are also careful to invert the answers for question 4, since this question is a negative statement while all of the other statements are positive. For this scoring rubric, a score of 1.0 implies very positive attitudes while a score of 0 implies extremely negative ones. The professors average all of the student scores to determine a course average and then create a histogram of pre and post-test scores (Fig. 1). Hyatt's course average changes from 0.61 to 0.57 and Armstrong's remains constant at 0.61. They interpret this to mean that traditional lecture courses have a decidedly negative effect on student attitudes while peer teaching is at worst neutral. Additionally, the histogram of student scores indicates that Professor Hyatt's course resulted in ~30% of her students post-testing with very poor attitudes, below 0.5. Finally, the student evaluations seem to back up this interpretation, and even suggest that peer teaching may have a positive effect, with some of Professor Armstrong's students writing:

"This course was fun"; "This course was more interesting than other science courses I have taken"; "I liked the way we talked about everything."

Professor Armstrong is happy with these results and decides based on this assessment that he will use peer teachers as a major component of all of his courses in the future. He is already reorganizing the upper division sedimentology course he teaches to incorporate peer teaching. Professor Hyatt is not as confident that peer teachers would add significantly to her lec-

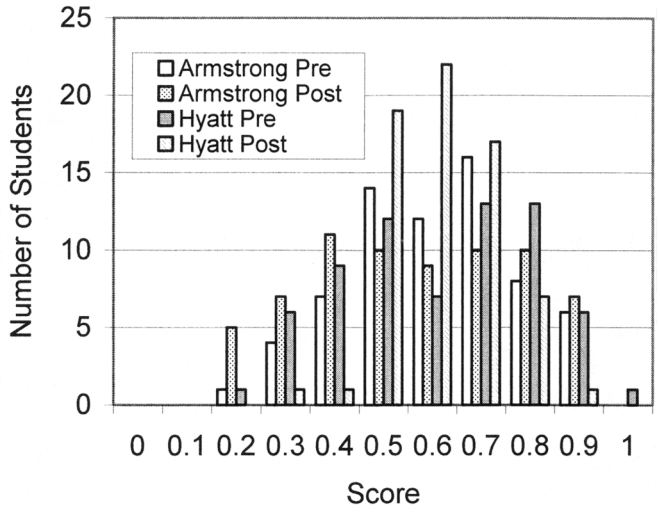

Figure 1. Attitude survey results from the courses taught by Professors Hyatt and Armstrong.

ture course, but she has decided to use peer instruction at least once next time she teaches an introductory course.

Do you see any of your own assessment efforts in the path Professor Armstrong has followed? How can his research design be improved to capture the most valid and reliable information?

THE INSTRUMENT

Professor Armstrong used an attitude survey published by other researchers. Because it was published, Armstrong thought, the survey must be a good means for gathering information on attitudes. However, it is very important to make sure that a published scale is valid and reliable before you use it for your own assessment purposes. Many authors conduct their own statistical analyses to ensure that their scales are indeed useful, although some authors do not. When deciding whether or not to use a published survey, whether you are testing content knowledge, skills acquisition, or another attribute, keep in mind that your study must be both valid and reliable.

What do the concepts of validity and reliability mean for science education research? Unlike the concrete physical or chemical phenomena observed in geology, the study of human responses will always be inexact. People, by nature, will never respond exactly the same way to the same external conditions. It would seem, then, that reliability and validity, the expectations that results are reproducible and accurate, can never be achieved. However, while it is true that people are much more varied in their response to stimuli than rocks or even plants, a wide array of research techniques have been developed by behavioral scientists to deal with this variability. The fields of education, psychology, and applied biology have a vast research literature documenting a variety of methodologies that we can use to evaluate curriculum, teaching methods, learning outcomes, and more. Additionally, validity and reliability can be documented with a variety of statistical techniques, including ANOVAs (analysis of variance), factor analysis, and item response theory. When using published instruments, ask yourself:

1) Does the author provide evidence that this test is valid? Validity implies that an expert population will score well on a test while a novice population will not. Choosing an expert population is not as simple as it may seem; expectations and realities are often not the same. Make sure you carefully consider who your expert group is, where they have been, and how the test relates to their own careers. For instance, although scientists are scientifically literate, they may not do well on a science content test outside of their field of expertise.

2) Does the author demonstrate the reliability of the test? Tests often consist of a series of multiple choice or Likert-scale questions. The test average is then used to determine a student's level of expertise. However, the internal consistency of the test itself can dramatically influence the meaning of the average score. That is, are the test items correlated with one another, such that a positive answer on one item implies a positive answer on all, and vice versa? Additionally, the test should produce the same results after multiple administrations.

3) Is it possible for you to test the reliability and validity of the test you wish to use? This may entail testing an "expert" population, such as science faculty or graduate students, to ensure that this admittedly scientifically literate population scores well on the test. You may also want to ensure that test items are linearly correlated with each other through a simple item analysis (Thorndike, 1997). A more complex factor analysis (Myers and Well, 1995) may ultimately be required to ensure that the scale is internally consistent. Additionally, more complex analyses, such as item response theory (Thorndike, 1997), can help you determine if the items are linearly correlated. However, an item analysis by itself can tell you a great deal. We can use the survey administered by Professors Armstrong and Hyatt as an example.

If Professors Armstrong and Hyatt had looked at the correlation between student responses on their survey, they would have found that only three of the items, questions 2, 4, and 5, are in fact correlated (Table 1). This is significant, as two-fifths of the test average used to determine student attitudes is the result of uncorrelated questions. That is, whether or not a student agrees or disagrees with statements 1 or 3 will have little to do with their overall attitudes towards science. Item analysis cannot give us any information about the validity (are we really testing attitudes?) of the test, but it is a very useful reliability indicator. You can perform a correlation analysis with a number of different computer programs.

If questions 1 and 3 seem to be assessing attitude, why don't they correlate to the other items? Questions 1 and 2

Question #	*1*	*2*	*3*	*4*	*5*	*Test Average*
1	1.0					0.39
2	0.08	1.0				**0.51**
3	-0.11	0.07	1.0			0.32
4	0.12	**-0.55**	0.12	1.0		**-0.58**
5	0.07	**0.68**	0.07	**-0.81**	1.0	**0.65**

Table 1. Item analysis of survey items used by Professors Hyatt and Armstrong based on 138 student responses. Ratings of 1.0 and 0 indicate perfect correlation and no correlation, respectively. Correlations in bold are significant; the negative correlations for question four reflect its negative wording. This analysis indicates that questions one and three are not correlated to the other test items or the final score.

actually contain learning style components; that is, 1 is assessing kinesthetic and 2, verbal, learning styles. However, it is quite possible to have positive attitudes about science and be ambivalent towards or not care for experimental work, rendering question 1 a poor item for assessing attitude. Question 2 may also not correlate for those individuals who do not like to read, hence the low correlation of 2 with other items, but within the tested population of college students this will be an issue for only a small subset of the students. The problem with question 3 is a little less obvious. This item implies that people who like science prefer it to all other subjects. Certainly, this is not always true. Additionally, even those students with positive attitudes have been known to dislike their science courses, so a preference for science classes over all others is not necessarily going to correlate well with attitude.

Interpreting the Survey Data. Although we have shown that all five test questions may not constitute the most effective test for determining student attitudes, Professors Armstrong and Hyatt still may be able to draw useful conclusions from the data. Primarily, the test average could be calculated using only the three test items, 2, 4, and 5, that are correlated with one another. Ideally, a test should be composed of as many items as possible, but since the professors already have this data in hand, they can garner some useful information from it. Additionally, instead of looking at the average course score and a histogram of individual scores, it can be more enlightening to look at the pre-test to post-test gain or loss experienced by each student (Figure 2, 3).

Professor Armstrong decided after comparing course averages and reading his course evaluations that peer teachers were a useful addition to his course. He has decided to use peer teachers in an upper division geology course. However, an analysis of pre-test to post-test changes indicates that sub-groups of students responded differently to his course (Figure 2). Primarily, those students who entered the course with poor attitudes, as reflected by an average pre-test score of 0.5 on the three correlated attitude items, experienced a positive change, or an increase in score on the post-test. Similarly, those students who entered the course with positive attitudes experienced a negative change. A statistical t-test or F-test can quantify the relationships demonstrated in Figure 2 by comparing average pre-test and post-test scores, as well as their distributions, for each subgroup. As with correlations, these statistical tests can be performed using a number of programs. For Professor Armstrong's course, those students pre-testing with negative attitudes improved from an average score of 0.38 to a post-test average of 0.55. So, the course did have a positive effect on this sub-group of students. However, students with positive attitudes experienced a decrease, from 0.70 to 0.58. This analysis indicates that using peer teachers may not be a good idea for those students who are already positive about science. Professor Armstrong should reconsider his use of peer teachers in an upper division course, although modification of the way in which they are used may alleviate the negative effect. This suggests another research project altogether.

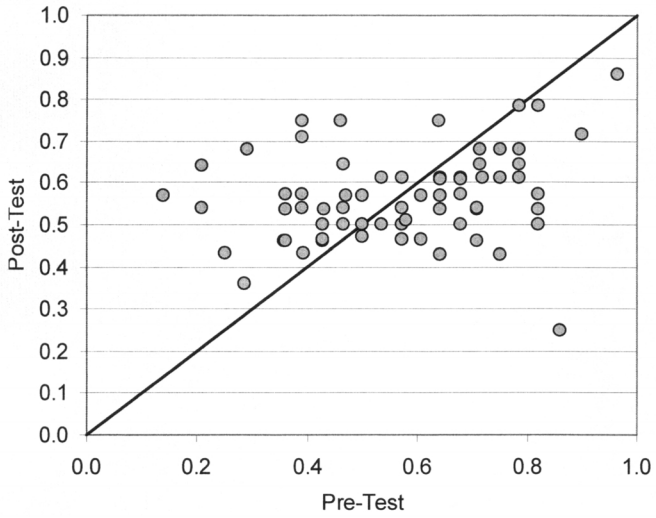

Figure 2. Professor Armstrong's course, collaborative learning component. Student averages for questions 2, 4, and 5 from the attitude survey. These results suggest that this course had a positive effect on those students who entered with poor attitudes (below 0.5), but had a negative effect on students who entered with positives attitudes (above 0.5).

Considering the data from Professor Hyatt's course reveals similarly interesting results (Figure 3). Students with poor initial attitudes experienced negative effects or no change as a result of taking the course, with an average decrease from 0.45 to 0.36 and statistical significance on the t-test. Interestingly, although it looks as if students with initially positive attitudes experienced an increase (Figure 3), a t-test indicates that the pre- and post-test scores are statistically identical. Professor Hyatt has already decided to use peer teachers at least once in her next

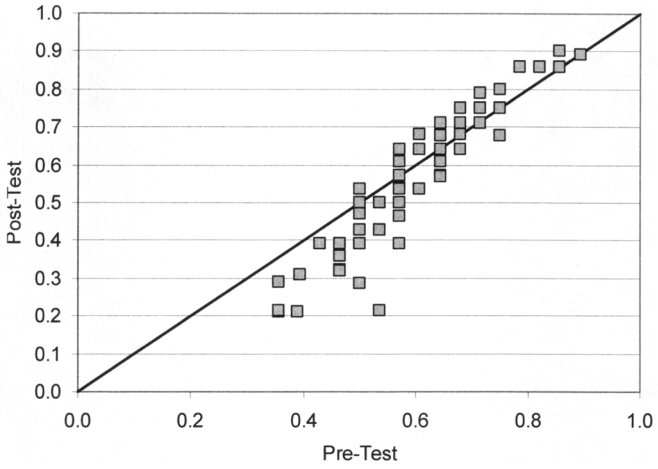

Figure 3. Professor Hyatt's course, primarily lecture-oriented. Student averages for questions 2, 4, and 5 from the attitude survey. Overall, this course had little effect on student attitudes. There may be a negative effect on those students who entered with poor attitudes, but the effect is not as pronounced as with Professor Armstrong's course.

course; based on the data from both classes, she will maximize positive effects if she chooses an activity that is targeted towards those students with the worst attitudes.

Professors Armstrong and Hyatt used student evaluations to provide anecdotal evidence of student satisfaction with the course, but they had no means by which to quantify the responses or correlate evaluations with the attitude survey. There are a number of established methodologies that can be used to evaluate qualitative data sets such as student evaluations (i.e. McTavish and Pirro, 1990). These methodologies will be discussed in a future column.

CONCLUSIONS

We have tried to provide a framework by which you can develop a research plan for assessment in your own classroom. The kind of data gathered and the statistical analyses employed will have a direct impact on any interpretations that can be drawn. Additionally, education literature deserves as much critical scrutiny as scientific articles. We hope that this column has begun to give you the tools you will need for your own research endeavors. Ultimately, you should be able to determine for yourself those teaching methodologies, curricula, and assessment tools which will be most useful to you as an educator, researcher, or both.

REFERENCES

Bair, E.S., 2000, Developing analytical and communication skills in a mock-trial course based on the famous Woburn, Massachusetts case: Journal of Geoscience Education, v. 48, p. 450-454.

Ewell, P.T., 1987, Establishing a campus-based assessment program, *in* Halpern, D.F., editor, Student outcomes assessment: What institutions stand to gain: San Francisco, Jossey-Bass, p. 9-24.

Johnson, A., 1997, Assessment, outcomes measurement and attrition: Reflections, definitions, and delineations: College and University, v. 73, p. 14-17.

McTavish, D.G., and Pirro, E., 1990, Contextual content analysis: Quality and Quantity, v. 24, p. 245-265.

Muehlberger, W.R. and Boyer, R.E., 1961, Space relations test as a measure of visualization ability: Journal of Geological Education, v. 9, p. 62-69.

Myers, J.L., and Well, A.D., 1995, Research design and statistical analysis: Mahwah, N.J., Lawrence Erlbaum Associates, 713 p.

Nitko, A.J., 1996, Educational assessment of students: Englewood Cliffs, NJ, Prentice-Hall, 482 p.

Shea, J.H., 1999, Education "research" at the annual meeting: Journal of Geoscience Education, v. 47, p.110

Terenzini, P.T., 1989, Assessment with open eyes: Pitfalls in studying student outcomes: Journal of Higher Education, v. 60, p. 644-664.

Thorndike, R.M., 1997, Measurement and Evaluation in Psychology and Education. 6th edition: Englewood Cliffs, NJ, Prentice-Hall, 583 pp.

Planning continuous improvement examines how an institution aligns what it wants or hopes to do with what it actually does. It examines an institution's systems and processes…

The Higher Learning Commission of the North Central Association of Colleges and Schools– Academic Quality Improvement Project, 2001