CBE—Life Sciences Education
Vol. 5, 313–316, Winter 2006

Letters to the Editor

Re: The Use of a Knowledge Survey as an Indicator of Student Learning in an Introductory Biology Course

Edward B. Nuhfer* and Delores Knipp†

*Department of Geology and Center for Teaching and Learning, Idaho State University, Pocatello, ID 83209; and †Department of Physics, United States Air Force Academy, CO 80840

Reliability is a fundamental quality of the internal consistence of any measuring instrument. Researchers sometimes use "reliable" and its derivative terms, yet fail to address reliability. Bowers *et al.* (2005) claimed that the knowledge survey (KS) "does not reliably measure student learning as measured by final grades or exam questions." They addressed their purpose (p. 311) "to evaluate how closely students' performance track with their confidence in their knowledge of the course material," through correlating "plotted pre- and post-KS scores against final grades (p. 314)." This approach assumes that tests/grades of unknown reliability are appropriate standards for judging other measures. Their article offers a case study in drawing conclusions without considering reliability.

The split-halves Spearman-Brown reliability (R) measure (Jacobs and Chase, 1992) derives from the r (r) obtained from individuals' scores on two halves of a single test. It is a routine method for quantifying reliability and is applicable to both tests and knowledge surveys. When Bowers *et al.* attributed specific claims to us: "They report that KS results represent changes in students' learning (p. 311)," and "... are a good representation of student knowledge (p. 316)," they omitted mention of the surprising reliability that characterizes knowledge surveys' pre- and postcourse measures (Figure 1).

Figure 2 shows a contrasting split-halves reliability for a faculty-made test of unusually high reliability that yields an r of r = 0.8 and a reliability coefficient of R = 0.9. Good faculty-made tests achieve R >0.6 (Jacobs and Chase, 1992), but many yield reliability coefficients of <0.3 (Raoul Arreola, personal communication, April 2005; Theall *et al.*, 2005). In correlating the test in Figure 2 with its equivalent KS, the maximum correlation to expect would be about r = 0.8 (the degree to which the less reliable instrument can correlate with itself). We believe r should be still lower, because we developed knowledge surveys to sample cognitive/affective domains that only partially overlap those sampled by tests (Wirth *et al.*, 2005).

Next, consider the reliability of grades derived by averaging several tests. An example appears in Figure 3. Suppose some tests lack significant correlation with other tests, but, as is common practice, we average all the tests anyway to produce grades. The reliability of grades derived from averaging a mix of tests should be lower than the most reliable tests used during the course. If we correlate grades and KS ratings, what, then, are we actually correlating? Without reliability measures, we cannot know.

Short-answer tests longer than 40 items usually have good reliability. Such tests define achievement based largely on knowing, as manifested through short-answer test-taking skills under timed conditions. Tests that address open-ended challenges sample other cognitive/affective domains to define achievement based largely on doing, as manifested through written reports and products generated after discussion, reflection, and revision. Such tests serve as much to promote learning and to mentor students to high-level thinking as to produce grades. High reliability is difficult to achieve using open-ended challenges, and quantifying it requires methods too cumbersome for routine classroom use. Many such tests within Figure 3 contribute to lower overall grade reliability than if achievement were derived solely from short-answer tests. One should know general reliability of one's tests/grades to understand what comparisons are possible, but it is less important to optimize a correlation between tests and knowledge surveys than it is to mentor students to engage open-ended challenges. Exceptionally effective courses may show near zero reliability in plots such as Figure 3. In such cases, paired comparisons are impossible, but the assessment value of a well-designed KS as a reliable, complementary measure becomes apparent.

Unrealistic expectations for high numerical correlation coefficients between grades and knowledge surveys persist until one understands limits imposed by reliability. Thereafter, understanding permits better interpretations. The positive numerical correlations Bowers *et al.* (2005) reported as "low" between their postcourse knowledge surveys and grades then seem surprisingly high, given limits imposed by reliability of tests and grades.

We emphasized assessment of learning in classes through use of aggregate data (Nuhfer and Knipp, 2003), which differs from Bowers *et al.*'s evaluative efforts to predict an individual's grades from her/his knowledge surveys. At the class level, we were most impressed by Bowers *et al.*'s Figure 1. It revealed the pattern change from no correlation be-

DOI: 10.1187/cbe.06–05–0166
Address correspondence to: Edward B. Nuhfer (nuhfed@isu.edu) or Delores Knipp (Delores.Knipp@usafa.af.mil).

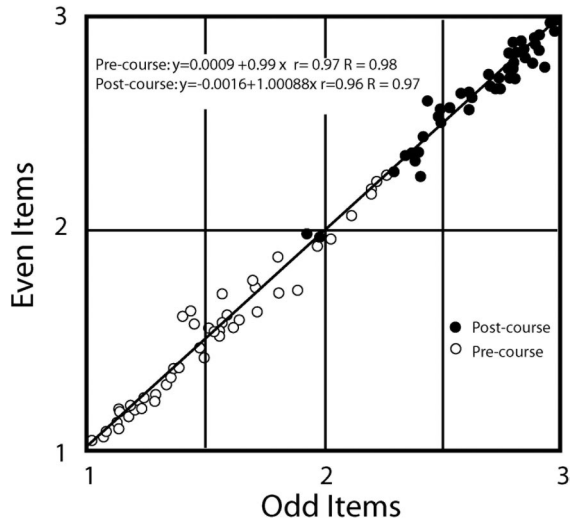

Figure 1. Scattergram of reliability (R = 0.98 precourse and 0.97 postcourse) measures on 49 students' pre- and postcourse knowledge surveys in an introductory geology course at Idaho State University. These students come from the same population as those described in Bowers *et al*.

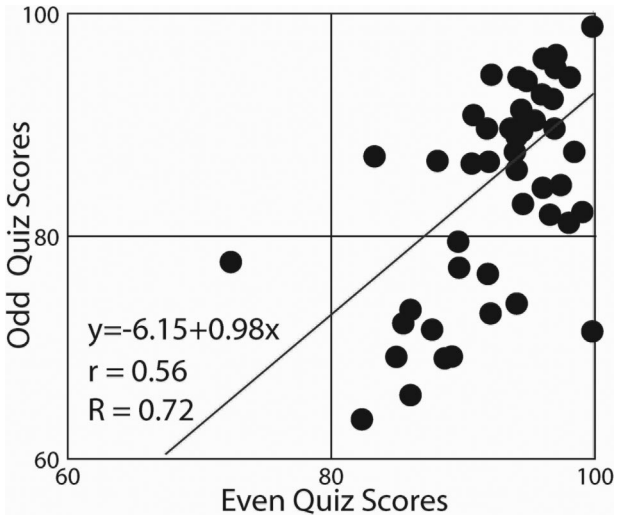

Figure 3. Scattergram of an overall reliability measure of course grades derived from 10 highly disparate quizzes (R = 0.72, for 48 introductory geology students at Idaho State University. The split-halves r of r = 0.56 reveals that, even if perfect relationships exist between the course knowledge survey and the grades, the highest r-value expected is only 0.56. The actual correlation was 0.47.

tween grades and the precourse KS to a persistent positive correlation between grades and the postcourse KS. Their tools (tests and the knowledge survey) remained constant through their pre- and postcorrelations, so the profound pattern change seen in every class seems most simply explained by students' increased understanding of specific content. We suggest correlating the high-reliability pre- and postcourse KS measures as an additional change indicator.

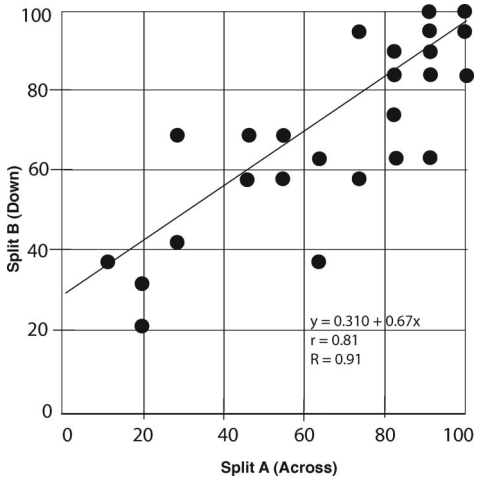

Figure 2. Scattergram of reliability (R = 0.91, derived from a split-halves r of r = 0.81) measures on a short-answer test (30 items, crossword fill-in-the-blank) in an introductory geology course taken by 25 students at Idaho State University.

Although Bowers *et al.*'s article derived from our work, neither the authors from Nuhfer's own Idaho State University campus nor the Journal's editors engaged us in review. The result is an article including several attributions to knowledge surveys and our thoughts/intentions that we disclaim. Readers who compare our ideas about knowledge surveys as presented by Bowers *et al.* with those published in our words (Nuhfer and Knipp, 2003; Theall *et al.*, 2005; Wirth *et al.* 2005) should anticipate discrepancies.

REFERENCES

Bowers, N., Brandon, M., and Hill, C. (2005). The use of a knowledge survey as an indicator of student learning in an introductory biology course. Cell Biol. Educ. *4*, 311–322. http://www.lifescied.org/cgi/content/full/4/4/311 (accessed 16 October 2006).

Jacobs, L. C., and Chase, C. I. (1992). Developing and Using Tests Effectively: A Guide for Faculty, San Francisco: Jossey-Bass.

Nuhfer, E. B., and Knipp, D. (2003). The knowledge survey: a tool for all reasons. To Improve the Academy. *21*, 50–78. http://www.isu.edu/ctl/facultydev/resources1.html (accessed 16 October 2006).

Theall, M., Abrami, P. C., Arreola, R., Franklin, J., Nuhfer, E., and Scriven, M. (2005). Valid Faculty Evaluation Data: Are There Any? AERA Annual Meetings Program Interactive Panel Presentation, American Educational Research Association Symposium, Montreal, April 14, 240. Summaries available at http://www.cedanet.com/meta/AERA2005valid.pdf (accessed 16 October 2006) and http://www.isu.edu/ctl/facultydev/extras/MeaningEvalsfract_files/MeaningEvalsfract.htm (accessed 16 October 2006).

Wirth, K. R., Perkins, D., and Nuhfer, E. B. (2005). Knowledge surveys: a tool for assessing learning, courses, and programs. Geological Society of America Annual Meetings Program with Abstracts. *37*, 7, 119. http://gsa.confex.com/gsa/2005AM/finalprogram/abstract_97119.htm (accessed 16 October 2006).

RESPONSE: Re: The Use of a Knowledge Survey as an Indicator of Student Learning in an Introductory Biology Course

Nancy Bowers* and Maureen Brandon†

*Instructional Design and Technology Solutions, Potsdam, NY 13676; and †Idaho State University, Pocatello, ID 83209

We appreciate Nuhfer and Knipp's comments on our article, and we acknowledge that there are multiple ways to analyze the effectiveness of knowledge surveys as they relate to student learning. Because of the paucity of peer-reviewed research concerning the implementation and interpretation of knowledge surveys, there is no generally accepted method for analysis. The statistical methods we used to address our question of the relationship between confidence levels (e.g., knowledge survey results) and student performance are sound and appropriate. We encourage others, including Nuhfer and Knipp, to publish their findings through the peer-review process, and we look forward to reading those articles.

DOI: 10.1187/cbe.06–07–0173
Address correspondence to: Maureen Brandon (branmaur@isu.edu).

RESPONSE: Re: The Use of a Knowledge Survey as an Indicator of Student Learning in an Introductory Biology Course

Diane Ebert-May and Everett P. Weber

Department of Plant Biology, Michigan State University, East Lansing, MI 48824

What is the correlation between students' confidence in their level of knowledge and comprehension in a course and their actual performance, as judged by their grades? The research by Bowers, Brandon, and Hill (Bowers *et al.*, 2005) investigated this question using knowledge surveys developed by Nuhfer and Knipp (2003) to determine students' perceived self-efficacy (belief in one's capability to carry out an action successfully). A knowledge survey (KS) is a series of content-based questions on topics presented in the course. Students are not asked to answer the questions, but merely to indicate for each question their *confidence* that they could answer it correctly. Evidence from Bowers *et al.* did not support the claim by Nuhfer and Knipp that students' learning can be predicted by perceived self-efficacy levels; rather, their data indicated that the correlation between student confidence and final grades is negligible. In response to this finding, Nuhfer and Knipp challenged the authors' interpretation of their results (see above letter). We are writing to comment on this controversy and to address a more fundamental question: How useful are knowledge surveys as assessment tools?

DOI: 10.1187/cbe.06–07–0174
Address correspondence to: Diane Ebert-May (ebertmay@msu.edu).

Designing assessments that provide substantive feedback about student learning in science presents a difficult challenge to faculty who teach undergraduates. The process of creating and evaluating assessments should include thinking broadly about validity and reliability. Bowers *et al.* focused on whether KS scores were valid measures of student understanding—*validity*. The authors concluded that the correlations they found between KS scores and student understanding were too low to validate a link. We concur with Bowers *et al.* that the statistical methods used in their study were appropriate and that the evidence supported their conclusions. However, Bowers *et al.* did not address the *reliability* of their assessments—that is, the reproducibility of the scores that would be obtained if the survey were administered several times to the same students.

Nuhfer and Knipp claimed high *internal reliability* of their instrument because the scores on related questions within the KS correlated highly with each other. In effect, this implied overall reliability of the instrument. They also suggested that the assessments used by Bowers *et al.* likely had low reliability, resulting in the observed low correlations between these assessments and the KS. Because of this factor, Nuhfer and Knipp argue in their letter that there may

indeed be a valid link between the KS scores and student understanding reported by Bowers *et al.*

Reliability and validity are both critical aspects of assessment. Differences in interpretation of assessment results occur and merit productive debate. As Nuhfer and Knipp point out, assessments that define achievement based on lower-level thinking (e.g., knowledge and comprehension) usually have high reliability. If the goal of instruction is for students to demonstrate gains in their knowledge and comprehension of the subject, the KS may be useful. However, science involves more than mastering facts. The KS is not designed to probe students' confidence about their ability to actively engage in processes of science or higher-level thinking such as analysis or synthesis. As we attempt to assess students' critical thinking abilities using open-ended problems, determining reliability and validity becomes more difficult (Batzli *et al.*, 2006). Ultimately, we wish to ascertain if assessments reliably measure what we want students to know and be able to do based on the goals and objectives of instruction. Are students becoming more sophisticated in their ability to solve complex problems that do not have single answers, high degrees of completeness, certainty or correctness? Is our instruction providing students guidance and practice in doing so?

Both papers agree that perceived self-efficacy is a key attribute when learning difficult subjects and addressing complex tasks. According to theory, as self-efficacy increases, students are more willing to undertake more complex tasks and think more about complex ideas and problems (Baldwin *et al.*, 1999). Instruction that nurtures critical thinking enables students to gain deeper understanding of content knowledge. Specific guiding questions can facilitate students' engagement with a problem and draw upon their prior knowledge or identify what they do not know (Ebert-May *et al.*, 2006). Pretests that identify students' comprehension or misconceptions about a topic can be used to guide instruction.

The peer-reviewed literature about knowledge surveys is sparse, and the relationship of perceived self-efficacy to performance merits further research. However, if critical thinking is the ultimate goal, a KS is unlikely to be a useful assessment. The potential value of knowledge surveys thus revolves around the question of whether "covering" content by the instructor is more important in an undergraduate science course than students "uncovering" content through problem solving and critical thinking. Whether instructors choose to use the KS should depend on their student learning goals. Perhaps instructors need to increase self-confidence in *their own* ability to promote higher-level thinking by their students.

REFERENCES

Baldwin, J., Ebert-May, D., and Burns, D. (1999). The development of a college biology self-efficacy instrument for non-majors. Sci. Educ. *83*, 397–408.

Batzli, J. M., Ebert-May, D., and Hodder, J. (2006). Bridging the pathway from instruction to research. Front. Ecol. Environ. *4*, 105–107.

Bowers, N., Brandon, M., and Hill, C. (2005). The use of a knowledge survey as an indicator of student learning in an introductory biology course. Cell Biol. Educ. *4*, 311–322. http://www.lifescied.org/cgi/content/full/4/4/311 (accessed 16 October 2006).

Ebert-May, D., Batlzi, J., and Weber, R. (2006). Designing research to investigate student learning. Front. Ecol. Environ. *4*, 218–219.

Nuhfer, E. B., and Knipp, D. (2003). The knowledge survey: a tool for all reasons. To Improve the Academy. *21*, 50–78. http://www.isu.edu/ctl/facultydev/resources1.html (accessed 16 October 2006).