Tinkering With The Traditional To Assess And Promote Quality Instruction: Learning From A New And Unimproved Instructor Evaluation Instrument

Audrey Amrein-Beardsley, Ph.D., Arizona State University, USA
Thomas Haladyna, Ph.D., Arizona State University, USA

ABSTRACT

For over 30 years survey instruments have been used in colleges of higher education to measure instructional effectiveness. Extensive research has been conducted to determine which items best capture this construct. This research study was triggered by a college of education's enthusiastic but failed attempt to create a new and improved instructor survey based on this research. Researchers found that the new instrument was no better than its predecessor. Student halo ratings contaminated results, reliability was lower than expected, and the survey results indicated one single dimension – general teaching effectiveness. Two associated variables of considerable interest, course relevance and rigor/demand, were also contaminated by student halo rating. Based on these findings and the extensive literature on student surveys of teaching effectiveness, we argue that traditional surveys based on conventional items may be valid for evaluating global teaching effectiveness and other summative purposes but not for the formative, self-diagnostic, and reflective purposes anticipated. New ways of evaluating teaching in higher education are presented and discussed. The article shares insights into theory-based survey development and a plan for validation.

Keywords: Course Evaluation, Instructional Effectiveness, Measurement, Reliability, Validity, Summative Assessment, Formative Assessment

INTRODUCTION

If we do not get out sleepers, and forge rails, and devote days and nights to the work, but go to tinkering upon our lives to improve them, who will build railroads? And if railroads are not built, how shall we get to heaven in season? Henry David Thoreau, Walden (1854)

In one of Henry David Thoreau's most famous books, *Walden*, he details one year of his life living in a remote cabin near Walden Pond, Massachusetts, a cabin owned by his friend Ralph Waldo Emerson. Thoreau wanted to better understand society and removed himself from it in almost complete solitude for a fictionalized period of four seasons. It is within the chapter, *Where I Lived, and What I Lived For,* Thoreau wrote the words quoted at the beginning of this manuscript. Although, in general, Thoreau resisted and critiqued development, progress, and the general tenets of modernization, he rejected complacency. With this railroad analogy, he entices us to question whether we are to tinker with the traditional and stationary to make false levels of progress or push forward to genuinely grow and evolve.

For the purposes of the research reported here, it is in higher education that his railroad analogy is translated and used to help us (1) understand the traditional process of evaluating instructor effectiveness and (2) move beyond tinkering with the inert and conventional processes to which we are accustomed in doing so. The practice of evaluating teaching effectiveness is critical to advancing the profession of teaching, in theory, but whether the traditional survey items used and with which researchers in higher education have tinkered for over 30 years, actually work in the ways purported and desired should be scrutinized and re-examined.

This particular re-examination was triggered by a college of education's enthusiastic but failed attempt to create a new and improved instructor evaluation survey. College faculty members and researchers believed that developing and constructing a new, research-based instructor evaluation survey would help college administrators better evaluate instructors' teaching and help faculty members use their students' responses in more formative, self-diagnostic and reflective ways. But, unfortunately for those who spent two years of good faith attempts to make the survey instrument the field standard (and fortunately for those who welcomed the disappointing findings as another opportunity for enriched understanding), it was determined that tinkering or experimenting with the repair of this, yet another traditional instructor evaluation instrument, was not an adequate solution to the problems they continued to face.

REVIEW OF RELEVANT RESEARCH

The use of a student survey of instructor effectiveness is and always has been the most commonly used practice mandated in colleges and universities, predominantly in Australia, Europe, and the USA (Al-Issa & Sulieman, 2007; Davies, Hirschberg, Lye, Johnston, & McDonald, 2005; Davies, Hirschberg, Lye, Johnston, & McDonald, 2007; Neumann, 2000; Richardson, Slater, & Wilson, 2007; Rowley, 2003), to (1) help administrators evaluate faculty members in terms of their instruction, (2) provide faculty with diagnostic feedback regarding how they can improve their teaching, (3) ensure high standards of teaching, and (4) ultimately improve instruction and student learning over time (Aleamoni, 1981; Dommeyer, Baum, Hanna & Chapman, 2004; Hendry, Cumming, Lyon, & Gordon, 2001; Johnson, 2000; McKeachie, 1997; Ngware, 2005; Roberts, Irani, Telg & Lundy, 2005).

The student survey of instructor effectiveness is usually used to determine things that capture the construct of "instructor effectiveness", including whether instructors are clear, knowledgeable, organized and prepared; are caring, courteous, enthused, friendly, helpful, professional and respectful; make course content practical and relevant; appropriately challenge students and hold them to high standards; encourage critical discussion, open communication and student interaction; are fair in terms of coursework, pace, and demands; are fair and timely in terms of grading coursework; and use data in reflective and formative ways (Carroll, 1963; D'Apollonia & Abrami, 1997; Feldman, 1976, 1988; Marsh, 1982; Nasser & Fresko, 2002). For a historical review of the research literature see Cohen (1981), Johnson (2000), and Ramsden (1991). For more detailed lists of traditional items used in such instruments see Feldman (1976, 1988), Li-Ping Tang (1997), Marsh (1982), Ramsden (1991), and Wilson and Lizzio (1997).

Overall, these survey instruments elicit reliable measures of student perception (Davies et al., 2005; Mohanty, Gretes, Flowers, Algozzine, & Spooner, 2005), and evidence has been presented to support the conclusion that these surveys effectively and defensibly measure the construct of instructor effectiveness (D'Apollonia & Abrami, 1997; Greenwald, 1997; Hellman, 1998; Hobson & Talbot, 2001; Johnson, 2000; Kogan & Shea, 2007; McKeachie, 1997; Nasser & Fresko, 2002). Some disagree (Ngware, 2005; Trout, 1997; Williams & Ceci, 1997). But because the results of these evaluations are used, usually in isolation of any other indicators of instructor quality to make promotion, tenure, and merit pay decisions; the survey instruments and the interpretations derived from their results are very controversial, even more so when instructors are compared to one another in normative, relativistic ways (Darby, 2006a; Johnson, 2000; McKeachie, 1997; Wilson & Lizzio, 1997).

When these surveys are designed and implemented at colleges and universities, their perceived levels of reliability and validity are widely taken for granted and go unquestioned. Validation is much needed, yet college and university personnel rarely undertake the effort needed to validate measures of instructional effectiveness (Al-Issa & Sulieman, 2007; Abrami, D'Apollonia & Cohen, 1990; Avery, Bryant, Mathios, Kang & Bell, 2006; Darby, 2007; Hobson & Talbot, 2001; Mohanty et al., 2005; Roberts et al., 2005). Instead, university personnel rely on the theory behind these instruments without investigating whether they can make valid inferences and high-stakes decisions from the results of these surveys (Greenwald & Gillmore, 1997; Haladyna & Hess, 1994; Ngware, 2005; Pounder, 2007).

A primary threat to validity of these student surveys comes from construct-irrelevant variance (CIV), a term used by Messick (1989) to describe factors that falsely inflate or deflate the measurement of a variable and therefore distort its interpretation. CIV is a threat to validity. A more commonly used term for CIV is *bias*, but this term is too general and does not capture specifically the kinds of variables that threaten validity (Haladyna & Downing, 2004). All measures of a construct, such as teaching effectiveness, have potential CIV, but too few researchers acknowledge these threats to validity, examine these measures for CIV, and if detected, implement the statistical controls needed to eliminate it (Cronbach, 1988; Haladyna & Hess, 1994; Kolitch & Dean, 1999; McKeachie, 1997; Moss & Hendry, 2002).

Haladyna and Hess (1994) suggested categories of *potential* threats to validity and are used here to organize current literature on this topic:

1. Student variables: Gender (as either interacts with the instructor’s gender), age, race and ethnicity, nationality, language background, year in school and undergraduate or graduate status, whether students perceive the course as relevant, content learned in the course, expected grade, received grade, prior achievement (GPA) and whether students perceive the evaluation as important;
2. Instructor variables: Gender (as either interacts with the student’s gender), age, race and ethnicity, nationality, language background, physical attractiveness, personality, charisma and other “edutainment” factors (Trout, 1997; see also Williams & Ceci, 1997), popularity of the instructor, academic rank and status, method of instruction, innovations and technologies used, whether a course is co-taught and grading leniency;
3. Course variables: Mode of delivery, class size, course time, course location (on- or off-campus), whether the course is required, the quantitative nature of the course (Davies et al., 2007) whether the course is part of a student’s major, course level, course workload, course content and perceived course difficulty; and
4. Instrument variables: Mode of delivery, response rates, the numerical scales used (Block, 1998; Gamliel & Davidovitz, 2005; Sedlmeier, 2006), perceived anonymity, an instructor’s influence on or “gaming of” students’ ratings (e.g. instructor presence or intentional/unintentional actions which positively or negatively influence student responses during form administration) and whether students believe instructors will use or benefit from their feedback to improve instruction.

Any of these sources of CIV may account for some variance in student ratings and depending on their collective effect, increase or decrease instructor ratings inappropriately. Thus, we should question the validity of the summative and formative decisions made based on instructors’ ratings where CIV is found to be present (Al-Issa & Sulieman, 2007; Aleamoni, 1981; Avery et al., 2006; Block, 1998; Cohen, 1981; Darby, 2006a, 2006b; Davies et al., 2007; Davies et al., 2005; Greenwald & Gillmore, 1997; Hellman, 1998; Hendry et al., 2001; Hobson & Talbot, 2001; Husbands, 1997; Marsh, 1982; Mohanty et al., 2005; Moss & Hendry, 2002; Nasser & Fresko, 2002; Oliver & Sautter, 2005; Pounder, 2007; Richardson, Slater, & Wilson, 2007; Ryan, & Harrison, 1995; Schmelkin, Spencer, & Gellman, 1997; Shevlin, 2000; Surridge, 2006; Timpson & Andrew, 1997; Trout, 1997; Williams & Ceci, 1997). Even open-ended, free-response items are threatened by CIV. Completion items added to a survey that require a written response are more commonly answered by females (Darby, 2006b). These sources of CIV should be considered or contextualized with logic and care, especially when high-stakes are attached to instructor ratings.

Further, whether students can make accurate judgments about the quality of their instructors is debatable (Aleamoni, 1981; Rowley, 2003; Wilson & Lizzio, 1997). Student ratings often display a halo effect (Hobson & Talbot, 2001; Schmelkin et al., 1997), which is defined as students rating the instructor holistically and responding to specific items with this holistic impression. This halo rating is present when students give the same holistic rating for each item instead of discriminating between items. An effective instructor would earn uniformly high ratings; an ineffective instructor would earn uniformly low ratings. And when students do this, the search for subscores that represent aspects of teaching is rendered impossible. The halo effect is a major theme in this research.

It is also questionable whether instructors can use results from these surveys to improve their instruction in formative ways (Aleamoni, 1981; Greenwald, 1997; Hobson & Talbot, 2001; Oliver & Sautter, 2005; Roberts et al., 2005; Wilson & Lizzio, 1997), particularly if useful sub-scores cannot be culled from survey instruments or open-ended responses are devalued or dismissed as subjective by those who perceive qualitative feedback in unconstructive ways (Johnson, 2000; Marsh, 1987; Trout, 1997).

BACKGROUND FOR THE DEVELOPMENT OF A NEW AND IMPROVED INSTRUCTOR EVALUATION INSTRUMENT

In 1986, a new college of teacher education adopted and developed a student survey to help them evaluate instructor effectiveness as directed by its governing board, the Board of Regents. A validation study was conducted 18 years later (Author(s), 2004) and indicated that there was considerable redundancy in the items due to student halo rating. Students were not discriminating among items, so there was very little inter-item fluctuation. Another finding was that subscores had approximately the same mean and were so highly correlated, that discriminative information from subscores seemed impossible. A factor analysis confirmed a finding of a single dimension. These items were so highly correlated that only one factor resulted – overall teacher effectiveness. In addition, instructor ratings were very high (a mean of 3.59 on a four point rating scale). There was a suspicion that students were displaying leniency in their ratings of instructors, yet the researchers also believed that teaching is excellent in the college because of the emphasis on teacher education. Consequently, the survey was limited in providing faculty with valid diagnostic information to improve teaching. It seemed reasonable to engage in substantial study and revision of the instrument.

In 2006, two years after its first validation study, the dean charged the college’s governance committee to develop a new, research-based instructor evaluation instrument to measure the construct of instructor effectiveness and replace the college’s traditional instrument. The governance committee appointed a committee of teacher evaluation, whose members decided to carry out the dean’s charge inclusively and invite all faculty members and key stakeholders to engage in the process of developing a new and improved instructor evaluation form.

Researchers noted that instructor evaluation instruments should be developed with administrative representatives who would ultimately use evaluation responses to make high-stakes decisions about faculty members; experts in testing, instrumentation, and validation research; instructional development experts; and faculty who would be most impacted by the official administration of the instrument (Aleamoni, 1981; Ngware, 2005; Timpson & Andrew, 1997). Organization support personnel were also involved as they were those who were to be in charge of the official administration, managing and storing of data, and reporting the results at aggregate and disaggregate levels.

Committee members reviewed the “most significant” research articles, as determined by resident faculty experts on the topic, written within the last 50 years. Members wrote summaries of each article and shared these summaries with all faculty members and key stakeholders in the college. Instruments from four other colleges on the campus and two other colleges within the university were also used to inform item development. These forms varied in terms of number of items included with a minimum of 10 to a maximum of 29 items, which the college’s survey instrument currently contained.

Two faculty forums were held to discuss and deliberate the general instructor evaluation research and the research on item development and refinement. Forty-three percent (20/47) of faculty members and stakeholders participated.

Participants decided to start with a larger set of items taken from the research pieces they deemed most useful. The predominant pieces of research came from the work of Feldman (1998) and Marsh (1982, 1997). Feldman (1998) analyzed items to identify exemplary teachers and correlated them with student achievement. He found that teacher’s preparation/organization of the course most highly correlated with higher grades in the course. Marsh (1982) developed the *Student Evaluation of Educational Quality* (SEEQ*)* instrument that captured nine significant factors which define effective teaching.

Ultimately the following eight items were pulled from these works to serve as the research-based core for the initial draft of the college’s new instrument:

1. Organization of the course
2. Teacher's clarity and student understanding
3. Whether the teacher pursued/met course objectives
4. Significance or perceived impact of instruction
5. Teacher's interest in the course/knowledge of the subject
6. Teacher's high standard of performance/motivation
7. Teacher's encouragement of questions/openness
8. Teacher's availability and helpfulness

Faculty participants also expressed four additional desires:

1. One factor identified by the faculty was the relevance of the course taught to one's future profession. If one teaches a course that students perceive as less relevant to the profession, does this unfairly affect one's rating of teaching effectiveness? Course relevance is construct-irrelevant because it isn't part of the definition of teaching effectiveness. Relevance is a belief by students that a specific course contributes to their professional competence. Some courses may have low relevance, regardless of who teaches the course or how effective that instructor might be.
2. Rigor does not seem to be a construct-relevant aspect of teaching effectiveness, but faculty members still wanted to capture whether what students learned was worth the cognitive demands of the course. Is rigor included in this definition of teaching effectiveness or is it a separate concept to be studied in relation to teaching effectiveness? If an instructor has more rigorous demands of students, how does it affect the instructor's overall ratings? The measure of rigor might be an additional variable to be considered when evaluating an instructor.
3. Based on the research, faculty members wanted to know what variables influenced mean course ratings and affected their overall scores. These variables included student's gender, instructor's gender, time of class, whether the class was held on- or off-campus, and students' expected grade. Since this was meant to be an instructor evaluation, items were to be centered on instructor effectiveness not course qualities. This category of variables is also construct-irrelevant.
4. It was also decided that overall comments would be solicited and open-ended responses would be solicited for every question within the instrument so students could choose to respond in more detail or better explain their scores by item, which would ultimately yield more formatively useful data.

In the spring of 2006, committee members developed the initial draft of the survey instrument. And throughout the spring, faculty members and stakeholders worked collectively to refine and eliminate items if deemed unimportant or unnecessary during full faculty meetings. They worked to add items to effectively capture rigor (e.g., "The cognitive demands were worth the time and effort I invested"). Some also expressed interest in having an item to capture the instructor's use of technology in the classroom. This item was included in the final instrument draft, yet this item was the source of much controversy, as expected.

In the fall of 2006, the final 15-item, revised, research-based survey instrument was pilot tested. This survey is presented in the appendix. A stratified random sample of students in 72 different classes were selected first by department and second by whether course instructors used BlackBoard, version 7.0, the university-sponsored, online course management system. Researchers wanted to ensure that the sample participating in the piloting of this instrument represented the four departments within the college and wanted to examine what effects, if any, piloting the new survey instrument online might have on student response rates. Up until this point, all students participated in these evaluations as long as they were present during class as members of the captive audience of students surveyed.

So in addition to completing the paper version of the current instructor evaluation, students selected to participate in this pilot were asked to evaluate their instructors using the new form administered online. Not surprisingly, the college response rate dropped from the high 90% range to 18%. Just 265 out of an approximate 1,440 students who were asked to complete the new instrument online participated. Students in classes in which instructors used BlackBoard were significantly more likely to participate in the online pilot than students not "online," although some online forms did not work for students who encountered issues with their internet browsers or computer cookies.

Because of the low response rate and its concomitant selection bias, faculty researchers determined the data were inadequate to make a sound decision about the validity of the new instrument. It also seemed some variables were sources of contamination. The two most important of these variables were course relevance and course demand/rigor, as hypothesized.

In the spring of 2007, an experiment was conducted to answer four research questions. These four questions comprise an important collection of evidence one uses to make a judgment about the validity of such survey instruments used for evaluating faculty members' teaching and providing diagnostic information to improve teaching. Below, questions are listed and the rationale for each:

1. What is the reliability of class/course means? Reliability is a necessary condition for validity for any measure of importance. Also, there is some conflict in the way reliability is estimated, particularly when student halo rating is present. Using coefficient alpha, if halo rating is present, these coefficients are spuriously high. Using parallel forms reliability provided a basis for estimating reliability without the threat of halo ratings.
2. What is the structure of the student responses? Supporting a single dimension? Supporting subscore validity for more diagnostic purposes? Content-related validity evidence is also a primary source for validating an important measure like this one (Kane, 2006). A major issue in this study was the development of valid subscore information when most evidence was expected to point to a general instructional effectiveness dimension.
3. How do two CIV variables, course relevance and course rigor, interplay with teaching effectiveness? As discussed previously, relevance should be independent of instructor effectiveness, and course rigor should not be strongly related to measures of instructional effectiveness. If anything, faculty hypothesized that instructors who were most rigorous would have been most likely to have low student ratings (see also Nasser & Fresko, 2002).
4. What about other CIV variables? As previously mentioned, many variables can undermine validity, and it is important to investigate these threats to validity with the idea of dismissing each threat or dealing with it.

RESULTS

The traditional (current survey) and the new (research-based) survey were administered to all students taking courses in the spring 2007 semester. A random, split-half method was used for assigning the two forms to students. That is, in every class half of the students completed the old, traditional form and half completed the new, research-based form on paper. Paper delivery returned the overall college response rate to approximately 90%, within the range of response rates historically posted. This helped justify further investigation. It was determined student responses could be used to compare the survey forms and better determine if the new, research-based survey would serve the college better than the traditional.

Despite the best efforts to achieve a perfect split-half assignment, some faculty members failed to administer the traditional form and some failed to administer the research-based form. Table 1 shows the sample sizes, means, and standard deviations for the traditional and research based forms.

Table 1: Descriptive Statistics for Both Forms Administered Split Half in 162 Classes

	Number	Mean	S.D.	Range	Reliability Estimate
Old and Traditional (Form A)	204	3.54	0.48	1 to 4	0.99
New and Research-Based (Form B)	169	3.62	0.38	1 to 4	0.98

	Number	Mean	S.D.	Range	Reliability Estimate
Old and Traditional (Form A)	162	3.59	0.42	1 to 4	0.57
New and Research-Based (Form B)	162	3.58	0.38	1 to 4	

Although the traditional form was given to 204 classes, only 162 of these classes received both forms as intended. A bias was introduced in this sample whereby the mean of the traditional form was 0.05 standard deviations larger than the mean of the traditional form for the smaller sample. The mean for the research-based form was 0.04 smaller for the total sample when compared to the mean for the sample of 162. This discrepancy was controlled for by eliminating all classes in which only one form was administered leaving the total sample size at 162 total classes. Additional details of this study can be found in (Author(s), 2009).

Question 1: Reliability

Reliability was estimated using coefficient alpha, an internal consistency index that depends on the intercorrelation among ratings. Coefficient alpha was very high, 0.99 for the traditional form and 0.98 for the new form. However, these results may be spurious because as mentioned previously in this article, halo rating was very likely. As noted previously, this result may have been due to students holistically rating their instructor.

Another way to estimate reliability is through scores obtained on parallel forms. As the traditional form and the research-based form were administered to 162 classes, we had a basis for calculating parallel forms reliability. Using the class means for each form, the reliability coefficient was 0.57, which is very low. Class means of student ratings of instruction tend to have a high reliability (see Gillmore, Kane, and Naccarato, 1976). The lower coefficient seemed to be a more accurate estimate of reliability because no halo rating was involved, and the two forms putatively measured the same construct, teaching effectiveness.

Question 2: Dimensionality and Subscores.

Regarding content, the results from both forms were strongly unidimensional. The high alpha coefficients and the results of factor analysis of responses to both survey forms provided very strong evidence of a single factor: general instructor effectiveness. This result is consistent with past studies previously cited.

A notable exception to the tendency for unidimensionality is found in the work of Marsh (1991, 1997) with the SEEQ. The items on this student survey are grouped by the subscale each item represents. This strategy may overcome the tendency for students to halo rate instead of discriminate among the components of teaching intended by the use of these subscores.

And as mentioned with reliability, halo rating may have also explained the result of this study. That is, students used a general impression of their instructor's teaching effectiveness instead of responding to each item uniquely. This kind of rating defeats the objective of creating valid subscores for diagnostic purposes. Subscores are much needed for formative evaluation if instructors are to improve their teaching.

Given this overwhelming evidence for one dimension, general instructor effectiveness, a more refined analysis was initiated to see if subscore information could be extracted from these student ratings. One clue to support this attempt to create valid subscores was that the item means varied more than what was expected by chance. For each item, the overall standardized class mean was subtracted from each instructor's standardized class mean. This resulted in a standardized residual value very much like an effect size. It shows how much above or below the average a particular instructor is.

An analysis of these residual values revealed some very important findings. Instructors with middle or high ratings did not have meaningful subscores to report. Their subscore means were tightly bunched around the mean of all items in the survey. Instructors with lower relative ratings, however, had meaningful subscores. Their subscores varied considerably, which indicated strengths and weaknesses in teaching useful for diagnostic purposes. Thus, this subscore analysis gave some weak evidence for validity yielding useful results for instructors with the lowest relative scores.

Question 3. Course Relevance and Course Rigor.

Regarding course relevance and course demand/rigor, the results did not support expectations. The correlation of relevance with instructor rating was 0.851, and the correlation of rigor/demand with instructor rating was 0.923. Was this result due to student halo rating? Or is course relevance and demand/rigor highly related to instructional effectiveness? As with the reliability and content-related validity studies just reported, halo ratings may have caused these results; the most rigorous courses got very high instructor ratings without exception. It was expected by the majority of faculty members that those instructors with the greatest rigor in their courses would get lower, not higher, ratings. Yet, the opposite was true. However, correlations between course demand/rigor and overall instructor effectiveness were extremely high which, in consideration of the ever-present student halo effect, might have been better explained by the one-dimensionality of the new assessment.

Question 4. Other CIV Variables

Regarding other variables that might exhibit CIV, several variables were found to be statistically significant, but practical significance was very small. Gender had a small effect, which showed that women rated their instructors slightly higher than men rated the same instructors. No gender effect was observed for the instructor or between the instructor and student. For time of day, instructors of evening classes received slightly higher ratings. For location (on- or off-campus), the latter had a small advantage.

SUMMARY OF FINDINGS AND CONCLUSIONS

Five overall findings resulted from this study:

1. Reliability is difficult to estimate accurately due to perceived student halo rating. When we correlated both forms to estimate reliability, the correlation coefficient was very low. This result suggests that halo rating is indeed present.
2. The data from both surveys appears to reflect a single dimension: general teaching effectiveness.
3. Using an extraction technique based on residual values, subscores were identified. These subscores seemed valid for instructors who scored well below the average.
4. Two important variables, course relevance and course rigor/demand, were problematic to measure. The relation between course relevance and instructional effectiveness should not be high, but was very high. This result implausibly suggests that courses that are most relevant are also the best taught. This result also implausibly suggests that courses that are most difficult are also the best taught.
5. CIV variables in this study were not found to be importantly related to instructional effectiveness. However, investigating CIV variables should always be done to allay any fears that such variables may contaminate interpretations of instructional effectiveness. Faculty should be consulted each year to determine what variables might influence ratings of instructional effectiveness unfairly and these variables should be considered when such examinations occur.

IMPLICATIONS

If college personnel go to tinkering with traditional instructor evaluations and their time-honored items, how can we get to using instructor evaluations to promote instructor quality, our ultimate end? In this process to develop a new instructor evaluation instrument, faculty members and stakeholders, as advised by the researchers who conducted these studies, concluded that their worldview, largely based on the parallel research community's worldview of the evaluation of instructor effectiveness, needed to change.

Again as advised, faculty members and stakeholders decided that these surveys should be short. Shorter forms would help the college achieve greater validity and would help to diminish the ever-present halo effects occurring, again, because traditional items within traditional evaluation instruments all point to what D'Apollonia and Abrami (1997) call the *General Instruction Factor* (p. 1203) and Hellman (1998) calls the *General Instructional Skill* (p. 45). Although a single factor structure may be useful for summative purposes, this factor alone is all but useful if instructors are to use student evaluation results in formative ways. Meaningful, valid subscores are needed.

The inclusion of open-ended items is still important for formative purposes (Nasser & Fresko, 2002; Tricker, Rangecroft, & Long, 2005). Providing students opportunities to respond in writing about specific items and/or the overall strengths and weaknesses of an instructor will help colleges and faculty target and better address instructional deficiencies. Better yet is the use of an action research or "loop" approach (Gapp & Fisher, 2006) using mid-semester feedback, collected once or more, formally or informally, which provides instructors with a "dipstick" evaluation allowing them the opportunity to continuously improve their instruction throughout each semester (Hendry *et al*., 2001; Hobson & Talbot, 2001; Kogan & Shea, 2007; Williams & Ceci, 1997). But to get at more formative data, traditional items like "the instructor was organized" or "the instructor spoke clearly" might be replaced with nontraditional "super-items" which would consist of a smaller set of global items collectively valued most (see also D'Apollonia & Abrami, 1997; Sedlmeier, 2006).

For example, super-items might themselves capture more global indicators of instructor effectiveness. These might include how much students perceived the course as relevant and worthwhile their learning (rigor) and how effectively the instructor delivered instruction, facilitated interactions, and evaluated student learning (see also D'Apollonia & Abrami, 1997). Colleges might consider following a similar, democratic process modeled here when developing these "super-items," allowing all faculty members and stakeholders, including students (Block, 1998; Cunningham & MacGregor, 2006), to determine what five to ten elements best define an effective instructor locally.

Colleges might also consider adopting a set of "super-items" that align with local and/or externally validated definitions of effective teaching. The National Board for Professional Teaching Standards (NBPTS) is probably the leading organization on this topic given its five core propositions of teaching effectiveness, otherwise known as what accomplished teachers should know and be able to do. Super-items aligned with these standards might capture the extent to which students perceive that (1) The instructor is committed to students and their learning; (2) The instructor knows the subjects (s)he teaches; (3) The instructor knows how to teach; (4) The instructor effectively manages and monitors student learning; and (5) The instructor is reflective (see http://www.nbpts.org).

There are other sets of standards (e.g. INTASC, NCATE, state teaching standards) that can be used to frame these instruments. But although the NBPTS model was developed for teachers who teach Pre-Kindergarten through twelfth grade, it is the only definition of the "teacher effectiveness" construct that has been validated by external researchers (Bond, Smith, Baker, & Hattie, 2000; Cavaluzzo, 2004; Goldhaber & Anthony, 2004; Vandevoort, Amrein-Beardsley, & Berliner, 2004) and translates easily into higher education settings. This model is becoming more universally accepted as colleges of teacher education across the country have begun to adopt and align these core propositions to help them evaluate and professionalize teaching in higher education across programs and processes (Cochran-Smith & Fries, 2001; Darling-Hammond, 2006; Dean, Lauer, & Urquhart, 2005; Schalock, Schalock, & Myton, 1998; Yinger, 1999).

An alternative option might be to adopt Chickering and Gamson's (1987) Seven Principles for Good Practice in Undergraduate Education which also promote high quality teaching and classroom learning as being active, learner-centered, and inquiry-oriented (see also Ngware, 2005). For these purposes it has been adopted by the American Association for Higher Education, the Education Commission of the States, and The Johnson Foundation (Chickering & Gamson, 1991). Super-items aligned with these standards might capture the extent to which students perceive that (1) Good practice encourages contacts between students and faculty; (2) Good practice develops reciprocity and cooperation among students; (3) Good practice uses active learning techniques; (4) Good practice gives prompt feedback; (5) Good practice emphasizes time on task; (6) Good practice communicates high expectations; and (7) Good practice respects diverse talents and ways of learning.

Either definition of effective teaching would work as a conceptual framework for the development of theory-based, new and improved instructor evaluation forms. Yet again, whether either option works effectively and in valid ways, as hypothesized, should be examined empirically at local levels.

CLOSING

In their book *Tinkering Toward Utopia,* Tyack and Cuban (1995) argue that, in general, educational innovations, reforms, and fads have ultimately failed primarily because educational reforms are constrained by and within highly homogeneous educational systems. Most educational reform as it exists today leans toward standardization and accountability. At the most, novel reform efforts have simply added to the existing organization temporarily, only to be debilitated and consumed by the larger, fixed system over time.

They argue that one the most constricting variables preventing educational reform success is, ironically, the teacher-centered classroom (see also Rury, 1997). So arguably, if we look to change the teacher-centered classroom in higher education and welcome student discovery, self-directed learning, inquiry-based classrooms, and constructivist modes of instruction and pedagogy, perhaps we should move beyond tinkering with that with which we are familiar and engineer railroads that may lead to greater student learning. In other words, we might use new and improved instructor evaluation instruments based on these “super principles” to help us and entice others to meet new ends (see also Nasser & Fresko, 2002).

At the same time, we need to remember that we as members of public institutions will be held and should hold ourselves accountable for that which we do. The need for a reliable and valid instructor survey is an integral, if not the most important piece, to help us satisfy this need to evaluate instruction. But unless the instrument with which we measure instructor quality challenges our fixed notions about education and drives that which we believe high quality instructors do, we will have yet one more meaningless, albeit temporarily exciting, failed attempt at substantive change.

AUTHOR INFORMATION

Audrey Amrein-Beardsley is an Assistant Professor at Arizona State University in the College of Teacher Education and Leadership. Her areas of research interest include teacher education, research methods, tests and assessments, and educational policy.

Thomas Haladyna is Professor Emeritus Arizona State University. During his career he has been an elementary teacher, teacher educator, test director, and research professor. He specializes in developing and validating testing programs and doing research on improvements in the measurement of educational constructs.

REFERENCES

1. Abrami, P. C., D’Apollonia, S., & Cohen, P., (1990). Validity of student ratings of instruction: What we know and what we no not. *Journal of Educational Psychology, 82*, 219-231.
2. Adamson, S. L., Banks, D., Burtch, M., Cox, F., Judson, E., Turley, J. B., Benford, R. & Lawson, A. E. (2003). Reformed undergraduate instruction and its subsequent impact on secondary school teaching practice and student achievement. *Journal of Research in Science Teaching, 40*(10), 939-957.
3. Al-Issa, A. & Sulieman, H. (2007). Student evaluations of teaching: Perceptions and biasing factors. *Quality Assurance in Education, 15*(3), 302-317.
4. Aleamoni, L. M. (1981). Student ratings of instruction. In J. Millman (Eds.), *Handbook of teacher evaluation* (pp. 110-145). Beverly Hills, CA: Sage Publications.
5. Author(s). (2004). *Validation of the student survey of instruction*. Reference to technical report to be added once peer-review process is complete.
6. Author(s). (under review). Validation of a research-based student survey of instruction in a college of education. *Research in Higher Education.*
7. Avery, R. J., Bryant, W. K., Mathios, Al., Kang, H., & Bell, D. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations? *Journal of Economic Education, 37*(1), 21-38.
8. Blackboard (1997). *Blackboard* (Version 7.0) [Computer software]. Washington, DC: Blackboard Inc.
9. Block, D. (1998). Exploring interpretations of questionnaire items. *System, 26*(3), 403-425.

10. Bond, L., Smith, T., Baker, W. K., & Hattie, J. A. (2000). *The certification system of the National Board for Professional Teaching Standards: A construct and consequential validity study*. Greensboro, NC: Center for Educational Research and Evaluation.

11. Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*(8), 723-733.

12. Cavaluzzo, L. (2004). *Do teachers with National Board Certification improve student outcomes?* The CNA Corporation: Alexandria, VA. Retrieved January 20, 2005, from: http://www.cna.org/documents/CavaluzzoStudy.pdf

13. Chickering, A. W., & Gamson, Z. F. (1987). Seven principles for good practice in undergraduate education. *AAHE Bulletin, 39*(7), 3-7.

14. Chickering, A. W., & Gamson, Z. F. (Eds). (1991). *Applying the seven principles for good practice in undergraduate education (New directions for teaching and learning, No. 47).* San Francisco: Jossey-Bass.

15. Cochran-Smith, M. & Fries, M. K. (2001). Sticks, stones, and ideology: The discourse of reform in teacher education. *Educational Researcher, 30*(8), 3-15.

16. Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*(3) 281-309.

17. Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

18. Cunningham, J. B. & MacGregor, J. N. (2006). The Echo approach in developing items for student evaluation of teaching performance. *Teaching of Psychology, 33*(2) 96-100.

19. D'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*(11), 1198-1208.

20. Darby, J. A. (2007). Evaluating course evaluations: The need to establish what is being measured. *Assessment and Evaluation in Higher Education 32*(4), 441-455.

21. Darby, J. A. (2006a). The effects of the elective or required status of courses on student evaluations. *Journal of Vocational Education and Training, 58*(1), 19-29.

22. Darby, J. A. (2006b). Evaluating courses: An examination of the impact of student gender. *Educational Studies, 32*(2), 187-199.

23. Darling-Hammond, L. (2006). Assessing teacher education: The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education*, (2), 120-138.

24. Davies, M., Hirschberg, J., Lye, J., Johnston, C., & McDonald, I. (2007). Systematic influences on teaching evaluations: The case for caution. *Australian Economic Papers, 46*(1), 18–38.

25. Davies, M., Hirschberg, J., Lye, J., Johnston, C., & McDonald, I. (2005). Is it your fault? Influences on student evaluations of teaching in tertiary institutions. Unpublished manuscript. Retrieved May 29, 2008 from http://tlu.ecom.unimelb.edu.au/papers/Is_it_your_fault.pdf

26. Dean, C., Lauer, P., & Urquhart, V. (2005). Outstanding teacher education programs: What do they have that the others don't? *Phi Delta Kappan,* (4), 284- 289.

27. Dommeyer, C. J, Baum, P., Hanna, R., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education, 29*(5), 611-623.

28. Feldman, K. A. (1998). Identifying exemplary teachers and teaching: Evidence from student ratings. In K. A. Feldman & M. B. Paulsen (Eds.), *Teaching and learning in the college classroom* (pp. 391-414). Needham Heights, MA: Association for the Study of Higher Education (ASHE) Reader Series.

29. Feldman, K. A. (1976). The superior college teacher from the students' view. *Research in Higher Education, 5*, 243-288.

30. Gamliel, E. & Davidovitz, L. (2005). Online versus traditional teaching evaluation: Mode can matter. *Assessment and Evaluation in Higher Education, 30*(6), 581-592.

31. Gapp, R. & Fisher, R. (2006). Achieving excellence through innovative approaches to student involvement in course evaluation within the tertiary education sector. *Quality Assurance in Education, 14*(2), 156-166.

32. Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1976). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement, 15*(1), 1-13.

33. Goldhaber, D. & Anthony, E. (2004). *Can teacher quality be effectively assessed?* Center on Reinventing Public Education: Seattle, WA. Retrieved January 20, 2005, from: http://www.crpe.org/workingpapers/pdf/NBPTSquality_report.pdf

34. Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, *52*(11), 1182-86.
35. Greenwald, A. G. & Gillmore, G. M. (1997). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89*(4), 743-51.
36. Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.
37. Haladyna, T. M., & Hess, R. K. (1994). The detection and correction of bias in student ratings of instruction*. Research in Higher Education, 35*(6), 669-687.
38. Hellman, C. M. (1998). Faculty evaluation by students: A comparison between full-time and adjunct faculty. *Journal of Applied Research in the Community College, 6*(1), 45-50.
39. Hendry,G. D., Cumming, R. G., Lyon, P. M., & Gordon, J. (2001). Student-centered course evaluation in a four-year, problem based medical programme: Issues in collection and management of feedback. *Assessment & Evaluation in Higher Education, 26*(4), 327-339.
40. Hobson, S. M. & Talbot, D. M. (2001). Understanding student evaluations: What all faculty should know. *College Teaching, 49*(1), 26-31.
41. Husbands, C. T. (1997).Variations in student evaluations of teachers' lecturing in different courses on which they lecture: A study at the London School of Economics and Political Science. *Higher Education, 33*(1), 51-70.
42. Kane, M. T. (2006). Content-related validity evidence. In S. M. Downing & T. M. Haladyna (Eds.) *Handbook of test development*, pp. 131-154. Mahwah, NJ: Lawrence Erlbaum Associates.
43. Johnson, R. (2000). The authority of the student evaluation questionnaire. *Teaching in Higher Education, 5*(4), 419-434
44. Kogan, J. R. & Shea, J. A. (2007). Course evaluation in medical education. *Teaching and Teacher Education: An International Journal of Research and Studies, 23*(3), 251-264.
45. Kolitch, E., & Dean, A. V. (1999). Student ratings of instruction in the USA: Hidden assumptions and missing conceptions about "good" teaching. *Studies in Higher Education, 24*(1), 27-42.
46. Lawson, A. E. (2003). *Using the RTOP to evaluate reformed science and mathematics instruction: Improving undergraduate instruction in science, technology, engineering, and mathematics.* Committee on Undergraduate Science Education. Washington D.C.: National Academies Press, National Research Council.
47. Lawson, A.E., Benford, R., Bloom, I., Carlson, M.P., Falconer, K., Hestenes, D., Judson, E., Piburn, M.D., Sawada, D., Turley, J., & Wyckoff, S. (2002). Evaluating college science and mathematics instruction. *Journal of College Science Teaching, 36*, 388–393.
48. Li-Ping Tang, T. (1997). Teaching evaluation at a public institution of higher education: Factors related to the overall teaching effectiveness. *Public Personal Management, 26*, 1997.
49. Marsh, H. W. (1982) Validity of students' evaluation of college teaching: A multi-trait, multi-method analysis. *Journal of Educational Psychology, 74*(2), 264-279.
50. Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, 11(3), 253-388.
51. Marsh, H. W. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, 83(2), 285-96.
52. McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*(11), 1218-1225.
53. Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement, 3rd ed*. (pp. 13-103.) New York: American Council on Education and Macmillan.
54. Mohanty, G., Gretes, J., Flowers, C., Algozzine, B., & Spooner, F. (2005). Multi-method evaluation of instruction in engineering classes*. Journal of Personnel Evaluation in Education, 18*(2), 139-151.
55. Moss, J. & Hendry, G. (2002). Use of electronic surveys in course evaluation. *British Journal of Educational Technology, 33*(5), 583-592.
56. Nasser, F. & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education, 27*(2), 187-198.
57. Neumann, R. (2000). A decade of student evaluation of teaching: Case study of an evolving system. *Journal of Institutional Research, 10*(1), 96-111
58. Ngware, M. W. (2005). An improvement in instructional quality: can evaluation of teaching effectiveness make a difference? *Quality Assurance in Education, 13*(3), 183-201.

59. Oliver, R. L. & Sautter, E. P. (2005). Using course management systems to enhance the value of student evaluations of teaching. *Journal of Education for Business*, *80*(4), 231-234.

60. Piburn, M., Sawada, D., Turley, J., Falconer, K., Benford, R., Bloom, I., & Judson, E. (2000). *Reformed Teaching Observation Protocol (RTOP) reference manual. ACEPT Technical Report No. IN00-3*. Tempe, AZ: Arizona Board of Regents.

61. Pounder, J. S. (2007). Is student evaluation of teaching worthwhile? *Quality Assurance in Education*, *15*(2), 178-191.

62. Ramsden, P. (1991). A performance indicator of teaching quality in higher education: the Course Experience Questionnaire. *Studies in Higher Education, 16*(2), 129-150.

63. Richardson, J. T. E., Slater, J. B., & Wilson, J. (2007). The National Student Survey: Development, Findings and Implications. *Studies in Higher Education*, *32*(5), 557-580.

64. Roberts, T. G., Irani, T. A., Telg, R. W., & Lundy, L. K. (2005). The development of an instrument to evaluate distance education courses using student attitudes. *American Journal of Distance Education, 19*(1), 51-64.

65. Rowley, J. (2003). Designing student feedback questionnaires. *Quality Assurance in Education, 11*(3), 142-149.

66. Rury, J. L. (1997). A book review of: *Tinkering Toward Utopia: A Century of Public School Reform. Teachers' College Record*. Retrieved December 3, 2007 from http://www.tcrecord.org/Content.asp?ContentID=9652

67. Ryan, J. M., & Harrison, P. D. (1995). The relationship between individual instructional characteristics and the overall assessment of teaching effectiveness across different instructional contexts. *Research in Higher Education, 36*(5), 577-594.

68. Schalock, D., Schalock, M., & Myton, D. (1998). Effectiveness (along with quality) should be the focus. *Phi Delta Kappan*, 468–470.

69. Schmelkin, L. P., Spencer, K. J., & Gellman, E. S. (1997). Faculty perspectives on course and teacher evaluations. *Research in Higher Education*, *38*(5), 575-592.

70. Sedlmeier, P. (2006). The role of scales in student ratings. *Learning and Instruction*, *16*(5), 401-415.

71. Shevlin, M. (2000). The validity of student evaluation of teaching in higher education: love me, love my lectures? *Assessment & Evaluation in Higher Education, 25*(4), 397-405.

72. Stone, S. L. & Qualters, D. M. (1998). Course-based assessment: Implementing outcome assessment in medical education. *Academic Medicine, 73*(4), 397-401.

73. Surridge, P. (2006). *The National Student Survey 2005: Summary Report*. Study sponsored by the Higher Education Funding Council for England (HEFCE).

74. Thoreau, H. D. (1854). *Walden*. Boston, MA: Ticknor and Fields.

75. Timpson, W. W., & Andrew, D. (1997). Rethinking student evaluations and the improvement of teaching: Instruments for change at the University of Queensland. *Studies in Higher Education, 22*(1), 55-65.

76. Tricker, T., Rangecroft, M., & Long, P. (2005). Bridging the gap: An alternative tool for course evaluation. Case Study. *Open Learning, 20*(2), 185-192.

77. Trout, P. A. (1997). What the numbers mean: Providing a context for numerical student evaluations of courses. *Change*, *29*(5), 24-30.

78. Tyack, D., & Cuban, L. (1995). *Tinkering Toward Utopia: A Century of Public School Reform.* Cambridge, MA: Harvard University Press.

79. Vandevoort, L. G., Amrein-Beardsley, A. & Berliner, D. C. (2004, September 8). National Board Certified teachers and their students' achievement. *Education Policy Analysis Archives, 12*(46). Retrieved January 20, 2005, from http://epaa.asu.edu/epaa/v12n46/

80. Williams, W. M. & Ceci, S. J. (1997). "How'm I doing?" Problems with student ratings of instructors and courses. *Change*, *29*(5), 12-23.

81. Wilson, K. L., & Lizzio, A. (1997). The development, validation and application of the Course Experience Questionnaire. *Studies in Higher Education, 22*(1), 33-53.

82. Yinger, R. (1999). The role of standards in teaching and teacher education. In G. Griffin (Ed.), *The education of teachers: Ninety-eighth yearbook of the National Society for the Study of Education* (pp. 85–113). Chicago, IL: University of Chicago Press.

APPENDIX

Instructor Evaluation Draft (Form B)

Please indicate your level of agreement with each of the following items. If you would like to write a comment for items 1-15 or any overall comments, please do so in the text boxes provided after each question. (Note: Textboxes have been removed from the appendix to save space).

Course Line Number: _______________	Strongly Agree	Agree	Disagree	Strongly Disagree
1. The coursework given by the instructor targeted course objectives.	□	□	□	□
2. The instructor was prepared to teach each class session.	□	□	□	□
3. The instructor made the course content clear.	□	□	□	□
4. The instructor demonstrated knowledge of course content.	□	□	□	□
5. The instructor contributed to my knowledge of course content.	□	□	□	□
6. The instructor showed interest in course content.	□	□	□	□
7. The instructor motivated me to learn course content.	□	□	□	□
8. The instructor provided opportunities for all students to participate.	□	□	□	□
9. The instructor was available to help students.	□	□	□	□
10. The instructor demonstrated respect towards students.	□	□	□	□
11. My learning was enhanced through the use of varied instructional techniques and tools.	□	□	□	□
12. The instructor held students to a high standard of performance.	□	□	□	□
13. The coursework was intellectually challenging.	□	□	□	□
14. The instructor put in place support systems to allow me to reach higher levels of learning.	□	□	□	□
15. I feel prepared to apply course content in my profession.	□	□	□	□

Please answer these background questions.

a. Was this a required course?	Yes □		No □		
b. In your opinion, how relevant was this course to your professional development?	Very relevant □	Relevant □	Not very relevant □	Not at all relevant □	
c. What grade do you expect to get in this course?	A □	B □	C □	D □	E □
d. How would you grade the amount of effort you exerted for this grade?	A □	B □	C □	D □	E □
e. What is your gender?	Male □		Female □		
f. What is the instructor's gender?	Male □		Female □		
g. At what time is this course is offered? *If the course is hybrid, at what time is the face-to-face class offered?	Morning □	Afternoon □	Evening □		
h. Where is the course offered	On-campus □		Off-campus □		

i. What is your major?		
	□ Bilingual Education □ Early Childhood Education □ Elementary Education □ Secondary Education □ Special Education	□ Master of Education – Administration □ Master of Education – Curriculum □ Master of Education – Special Education □ Doctor of Education □ Other

j. During face-to-face sessions, approximately what percent of instruction was lecture or activity based?				
□ 100% lecture	□ 75% lecture 25% activity	□ 50% lecture 50% activity	□ 25% lecture 75% activity	□ 100% activity

k. Approximately what was the ratio of course sessions which were conducted face-to-face (f2f) versus online?				
□ 100% online	□ 75% online 25% f2f	□ 50% online 50% f2f	□ 25% online 75% f2f	□ 100% f2f

Please write in your comments about this course in the box provided.