International Journal for the Scholarship of Teaching and Learning

Volume 5 | Number 1 Article 11

1-2011

Developing a Statistically Valid AND Practically Useful Student Evaluation Instrument

Jeffrey S. Skowronek
University of Tampa, jskowronek@ut.edu

Bruce K. Friesen
University of Tampa, bfriesen@ut.edu

Heather Masonjones
University of Tampa, hmasonjones@ut.edu

Developing a Statistically Valid AND Practically Useful Student Evaluation Instrument

Abstract
The current article presents the findings on the development of a student evaluation instrument in which course evaluation is directly tied to student learning outcomes. With a committee consisting of instructors from six distinct disciplines brought together as part of a working group for this purpose, the instrument was developed utilizing research on the components of effective teaching and how these components impacted student learning. The instrument was tested at two time points, once via pen and paper (n=340 students) and the other online (n=2636 students). Factor Analysis resulted in one latent factor both times. The instrument also had high internal consistency reliability. Comparisons of individual student factors revealed a few variables significantly predicted ratings, but effect sizes were small. This work suggests an instrument has been created that assesses components of effective teaching, via the impact on student learning, and the ratings obtained are not highly influenced by individual factors.

Keywords
Student evaluation instrument, Student learning outcomes

Developing a Statistically Valid AND Practically Useful Student Evaluation Instrument

Jeffrey Skowronek
jskowronek@ut.edu

Bruce Friesen
bfriesen@ut.edu

Heather Masonjones
University of Tampa
Tampa, Florida, USA
hmasonjones@ut.edu

Abstract

The current article presents the findings on the development of a student evaluation instrument in which course evaluation is directly tied to student learning outcomes. With a committee consisting of instructors from six distinct disciplines brought together as part of a working group for this purpose, the instrument was developed utilizing research on the components of effective teaching and how these components impacted student learning. The instrument was tested at two time points, once via pen and paper (n=340 students) and the other online (n=2636 students). Factor Analysis resulted in one latent factor both times. The instrument also had high internal consistency reliability. Comparisons of individual student factors revealed a few variables significantly predicted ratings, but effect sizes were small. This work suggests an instrument has been created that assesses components of effective teaching, via the impact on student learning, and the ratings obtained are not highly influenced by individual factors.

Introduction

Since the inception of student evaluation instruments in the 1960s (Cahn, 1986), there have been concerns about the reliability, validity, and appropriateness of these tools in assessing the quality of courses and professors. While detailing the controversy surrounding the issues in using student evaluations is beyond the scope of this paper, a few issues are worth noting. Questions of reliability currently appear to be resolved; student evaluations appear to be reliable both between ratings made by different students for the same course and for ratings made by the same student over time (Huemer, 2005; Marsh & Roche, 1997; McKeachie & Hofer, 2001).

Student evaluations appear to be somewhat valid, especially when compared to other indices of teaching effectiveness or student achievement (McKeachie, 1997; Ory and Ryan, 2001). At the same time, a number of assumed biases with student evaluations can affect ratings even though they are not directly related to course or teaching quality; these include: grade leniency and effects of course difficulty (Huemer, 2005; Trout, 2000), level of showpersonship or the "Dr. Fox Effect" (Marsh & Roche, 1997; Ware & Williams, 1975), differences between departments such that more science oriented disciplines receive lower ratings (Cashin, 1990; Basow & Montgomery, 2005), and subjective student factors such

https://doi.org/10.20429/ijsotl.2011.050111

as a search for personal meaning versus the acquisition of knowledge per se (Entwistle and Tait, 1990). Many agree that student evaluations can be an integral part of the evaluation of an instructor's performance (Marsh & Roche, 1997; McKeachie, 1997, McKeachie & Hofer, 2001; El Hassan, 2009). Methodologically, it is critically important to address issues relating to construct validity; answering "Does the nature of the student rating process fit the construct being measured?" Messick (1989) identifies six dimensions of construct validity as it pertains to student evaluations: content, substantive, structural, generalizability, external, and consequential. Of these, Ory and Ryan (2001) report a paucity of research in the areas of content, substantive, and consequential validity. In a Beirut study, El Hassan's (2009) research addresses issues of substantive and consequential validity, reporting that they can be meaningfully addressed in evaluation efforts that are well-planned and executed, including effectively communicating to students and faculty the purposes of the evaluation process. As of yet, relatively little attention has been paid to the issue of *content validity*; the extent to which a measure essentially captures a given social construct.

Richardson (2005) describes several important student evaluations of teaching (SET's) in use in research projects in the U.S.A., England, and Australia; most notably Marsh's (1982) Student Evaluations of Educational Quality (SEEQ), British student satisfaction surveys such as the Noel-Levitz Student Satisfaction Inventory, Ramsden and Entwistle's (1981) Course Perceptions Questionnaire (CPQ), and Ramsden's (1991) Course Experience Questionnaire (CEQ). Rigorously tested, such SET's are successful in that they appear to adequately measure what they attempt to measure: quality of teaching, student satisfaction, educational experience, or global evaluations of departmental or programmatic curriculum. Missing from this list is a more directed attempt to assess the extent to which these efforts are correlated with the ultimate point of education: the amount a student has learned. Aside from tying teaching efforts to desired outcomes, asking students questions about their learning reinforces the intent of classroom efforts and activities (Titus, 2008).

What follows is an account of a diverse group of teachers brought together as a working group to examine an existing SET used across the University. In rejecting outright the instrument in use, committee members reasoned through issues of content validity and usefulness as they worked to build a new student evaluation of learning (SEL). In reconstructing this narrative we lay bare the logic by which the new instrument was developed, thereby adding to the literature on the construct validity of such instruments.

Background

In 2003, a "teaching effectiveness task force" was created at a small-medium sized university in Florida to address the issues being faced with student evaluations, including the validity of the instrument, the appropriateness of the items, and the proper use of the ratings in the instructor's overall evaluation process. Although the student evaluation instrument was partially revised, the committee's work was left unfinished when it was dismantled due to other pressing university concerns.

In 2005, however, the committee was reestablished to examine the University's student evaluation of teaching. This committee was charged with three goals for the SET: 1) standardized in a way that would provide administrators with comparable information to use in decision making, 2) diagnostic information that would provide individual faculty with meaningful feedback to improve their teaching, and 3) adoption of an SET suitable for use in an online format. In achieving these goals, it was immediately clear to committee members that the current instrument needed to be replaced. The instrument in use was

made up of evaluative and subjective statements (e.g., the professor's *high* level of enthusiasm), was biased towards certain disciplines, and even appeared to leave certain applied, or more creative-oriented, disciplines in the University's "blind spot." As a result, the committee, made up of six members from six distinct disciplines (Art, Biology, Communications, Psychology, Sociology, and Theatre), began the process of trying to create what would ultimately be a statistically valid, but also practically meaningful, student evaluation instrument. Several committee members were well-versed in the literature on teaching excellence and scale development techniques.

Content Validity: Developing Instrument Items

The creation of the instrument began with lengthy discussions of what qualities were *essential* to be an "effective teacher" across all disciplines. These multidisciplinary considerations were based on experience and grounded in supporting research and literature. This proved to be a humbling experience, as essential components of quality teaching in one area (such as organization and quality readings in social science courses) were not necessarily critical elements of good teaching in another area (such as exploration and creativity in a sculpting class, for example). Such dialogue pushed committee members to identify truly universal elements of quality teaching. Gibbs (1995) argued that generating this definition was an essential first step in evaluating quality teaching. The inherent difficulty in defining effective teaching is obvious; effective teaching is a complex, dynamic issue that varies by subject matter and even personality (i.e., what works for one teacher may not work well for another). Furthering the difficulty was the belief that great teachers are "born, not made" (McKeachie and Hofer, 2001) and good teaching does not come with "technique" (Palmer, as cited in Baiocco and DeWaters, 1998). Whether or not great teaching ability is innate (Bain, 2004), in order to benefit from classroom evaluations there must be a belief that educators can at least learn to be *good and effective* teachers and that this learning can come from external feedback.

As a result of the multitude of issues, a clear definition of teaching effectiveness continues to elude educators (as evinced by the continued emergence of teaching metaphors relating excellence to "The Wizard of Oz" and Machiavelli's "The Prince" (Teverow, 2006)). For purposes of this study, a ***working definition*** was developed to include that an effective teacher: 1) creates an active learning environment to engage students (Angelo, 1993), 2) makes an attempt to identify students' prior knowledge about a topic and goals for a course (Perry, 1970), 3) attempts to make course content meaningful to the "real-world", 4) attempts to develop deep levels of understanding and help students reflect on that understanding (i.e., critical thinking) (Halpern, 1999), 5) should remain excited and enthusiastic about the material they are teaching (Voss & Gruber, 2006) and 6) is committed to personal growth within the discipline (Lowman, 1995). At its pinnacle a teacher must serve as the ultimate model of learning. While there may be other components that need to be added, this working definition was used as a building block to identify core qualities of the effective teacher.

An Innovation in Measurement

Once the core components of effective teaching were established, instrument items were generated. Along with the set of new assessment items, a new rating scale was created. This scale was adapted from a model used at the University of California-Berkley in assessing "student learning gains" (UC Regents, 2000). Rather than asking if students agreed or disagreed with a statement (on a five point scale ranging from strongly disagree to strongly agree), students are now being asked whether or not a certain component helped their learning (on a five point scale ranging from "did not help my learning" to

"helped my learning a great deal"). Learning was eventually defined as "a sustained and substantial change in the way a student thinks, acts, or feels" (Bain 2004).

This was a dramatic shift in the student evaluation instrument as focus was shifted from emotive responses regarding instructional methods to a focus on what the instructor does to facilitate learning (i.e., a student might not agree with the presentation style an instructor used, but he or she could still learn in such an environment). The following course characteristics were eventually selected for inclusion in the instrument: class structure, pace, assignments/projects/activities, and discussions. Instructor behaviors included: presentations, enthusiasm, stimulation of interest, student interactions, feedback, and challenge for self-betterment (see Appendix A). Even though questions selected were intended to be essential for all disciplines, a "not applicable" response was included in the instrument (Shuman & Presser, 1979).

Wording of instrument items was evaluated to ensure non-sexist, non-evaluative, and non-subjective language. To separate potential confounding or multidimensional issues of teaching, no questions assumed any quality or component was present in the classroom. Instead, additional items were added to allow students an opportunity to first provide information about the level of certain components. For example, students were asked as to what *level* of enthusiasm the instructor seemed to exhibit, and were then asked how the level of enthusiasm impacted *learning*. The level/learning distinction was the result of continued committee debate over whether or not the student evaluation instrument should assess the methods used in teaching or the outcomes of those methods (i.e., impact on learning). We believed that separating the presence of a characteristic from its impact on learning helped us avoid issues of multidimensionality present in other measures (see Ramsden and Entwistle, 1981; Marsh, 1991).

A deliberate effort was extended, then, to measure a unified concept: teaching efforts that necessarily produce quality outcomes. While impacting student learning in some positive way is the goal for all teachers, important techniques have been identified in the literature that can be used to maximize the possibility of learning. This was addressed in our working definition of teaching effectiveness. Most other SET's assess either method or outcome. We have seen few instruments that actually address both, as in this new instrument.

Finally, comment boxes were included immediately following many of the items for students to provide specific narrative feedback in addition to the more global narratives typically provided at the end of an evaluation survey.

Focus Group Assessment

Once the instrument was developed, a student focus group was conducted to get essential feedback on how each item was being interpreted and how the overall scale was viewed. This was believed to be a crucial step in assuring the validity of the instrument before it was piloted in the university. Twenty students of various college status and disciplines were chosen to participate in the focus group. The students were informed that they were evaluating a new "classroom survey" and should read each item closely. In order to get useful data and provide a focus while completing the instrument, students were asked to evaluate "the first professor that came to mind." Hoping to obtain the most honest results, all surveys were completed anonymously and the student did not identify the professor chosen.

Upon completion, students were asked open-ended questions about each question and the survey as a whole. Student responses were positive. Students appreciated how the survey focused on learning ("I could loathe the professor but still learn a lot"). They also liked having comment boxes after items to provided specific feedback. Students provided information about the order in which questions appear, wording, and interpretation of items (what words like "pushed" or "challenged" meant to the student). Even with this small sample size, the responses were found to have high internal consistency reliability, Cronbach's $\alpha = 0.94$, suggesting the pattern of results were similar for all students and the items were rating a similar latent quality.

Time 1: Pencil and Paper Testing- Fall 2007

Method

Participants

Twelve professors teaching 26 courses across nine disciplines at a small-medium sized private university in the Southeast United States were used in this testing. Five of the professors participating in the pilot were part of the committee that created the survey, while the remaining seven came from a group of professors that were asked to volunteer to participate. A total of 340 students anonymously completed the survey (98 males, 242 females) and were included in the analyses. Forty-three (12.6%) of the students were Freshmen, 83 (24.4%) were Sophomores, 101 (29.7%) were Juniors, 109 (32.1%) were Seniors, and 4 (1.2%) were either Graduate Students/Other or did not report the year in college. Two-hundred forty seven (72.6%) students who responded to the race/ethnicity question were Caucasian/White.

Procedure

Three weeks prior to the end of the semester, all university instructors agreeing to participate in the pilot study received packets containing copies of the instrument, now called the "classroom survey," and specific instructions for both the instructor and students. Instructors were asked to have the survey completed at the beginning of the class session and to allow approximately 20 minutes for completion. Prior to beginning the survey, students were informed that they were part of a pilot study and were using a newly developed instrument. As a result, the students were provided an overview of the new rating scale and were informed that they were to rate the impact of learning rather than how much they agreed with a statement. As with any course evaluation, a brief set of instructions were read to the student and the instructor left the room while students completed the survey. A student was asked to collect the completed surveys in a packet and, when all surveys were collected, return the sealed packet to the Dean's office.

The Instrument

The "classroom survey" was split into three sections: one section each pertaining to the course, the professor, and the student (See Appendix A). Five components were assessed in the course section, including: structure, pace, assignments, discussions, and exams. Seven components were assessed in the professor section, including: presentation quality, enthusiasm, stimulating interest, interaction with student, feedback provided, challenging students, and use of course readings. All questions that assessed an impact on learning were rated on a five-point scale where 1= "Did not help my learning," 3= "Helped my learning adequately," and 5 = "Helped my learning a great deal." In each section there were also some questions related to the level of certain qualities, including: pace, discussion, enthusiasm, stimulation, feedback, and challenge. Seventeen items were assessed in the

student section, including: gender, status, major/minor (answered yes/no) in department of course rating, prior courses in department, hours per week spent on class, percent of class sessions fully prepared for, and expected grade in the course (A, AB, B, BC, C, CD, D, F).

Results

Scale Construction

Factor Analysis was conducted to determine the underlying latent structure of all the items that assessed *impacts on learning*. Items related to "levels" of certain components were not analyzed because they are simply meant to be qualitative information for the professor. Principal axis factoring was conducted using a varimax rotation. Any factor with an eigenvalue over one was retained. In order to be included as part of the factor, items had to load .5 or higher (we used .5 as a very conservative value to ensure the items truly did relate to the latent factor). The resulting factor structure produced only one factor, labeled "teaching effectiveness." Because there was only one factor there was no need for rotation. This factor accounted for 56.64% of variance. All the items loaded positively (See Table 1).

Table 1. Factor Loadings of Survey Items Related to the Impact on Learning for Time #1

Question section (Course or Professor), Question number, and Question focus	**Loading**
Course #1- Structure	.801
Course #2b- Pace	.677
Course #3- Assignment/projects	.692
Course #4- Class discussion	.696
Course #5- Exams	.660
Professor #1- Presentations/explanation	.779
Professor #2b- Enthusiasm	.794
Professor #3b- Stimulate interest	.822
Professor #4- Interactions	.782
Professor #5b- Feedback/comments	.751
Professor #6b- Challenge	.803
Professor #7- Use of Readings	Removed

Only one item did not meet the criteria for inclusion in the factor, with a loading of .456, and was removed from further analysis. The level of internal consistency reliability was high, Cronbach's α =.9335. This was similar to the value obtained in the focus group study and could not be improved by deleting any of the items.

Individual Factors

After the factor analysis, all "learning" items were summed to create a "teaching effectiveness score." The responses to the student information, which is defined as any response not part of the course or professor rating, were then compared to the teaching effectiveness score in a correlational matrix. Eight of 14 items were significantly correlated to the teaching effectiveness score. In order to assess the unique contribution of each "individual factor" (Mauer, et al., 2006) a multiple regression predicting teaching

effectiveness was conducted. The overall model was significant, F (8,326) = 11.014, p =.0001, r^2 = .193.

Of the possible 8 individual factors that were thought to possibly influence ratings, six were significant predictors of teaching effectiveness: these factors included whether or not the student sought the professor's assistance, percent of time the student was fully prepared for class, how often the student participated in class discussions, and the expected grade for the course (β, p, and sr^2 values are presented in Table 2). Of these items, "believed grade" accounted for the most unique variance, sr^2 =.185.

Table 2. Time #1- β, p, sr^2 (squared semi-partial correlation) for multiple regression predicting teaching effectiveness score (Covariates: # of missed classes, hours spent per week on course, seeking professors assistance, % of time fully prepared for class, participation in discussion, % of assignment completed, % of course readings completed, and believed grade)

Student Variable	β	**p**	**sr^2**
Seeking professor assistance	.159	.003	.145
Being fully prepared for class	.196	.0001	.173
Participate in class discussion	.117	.037	.103
Believed grade	.204	.0001	.185

Time 2: Online Testing- Spring 2008

Method

Participants

One hundred and twelve professors teaching 276 courses across 17 disciplines volunteered to participate in Time 2 of the survey testing. This testing was part of a large- scale university attempt to move the course evaluation process on-line and participation was requested by a Dean to all faculty in the college. A total of 2636 students anonymously completed the survey (599 males, 2011 females, 26 gender unreported). Seven hundred and twenty-five (27.5%) of the students were Freshmen, 619 (23.5%) were Sophomores, 648 (24.6%) were Juniors, 577 (21.9%) were Seniors, and 67 (2.5%) were either Graduate Students/Other or did not report the year in college. A total of 1903 (72.2%) students reported they were Caucasian/White. Missing data reduced the total useable sample to 1814 students.

Procedure

One month prior to the end of the semester, all university instructors agreeing to participate in the Time 2 test received a detailed e-mail instructing them on how the evaluations would be administered to their classes. They received a copy of the "classroom survey" and specific instructions for both the instructor and students. Prior to beginning the survey, students were informed that they were part of the testing of both the new instrument and

also the process of completing the survey on-line. A private survey company notified each student by e-mail when the survey window opened and provided them with a unique password for each course survey they needed to complete. Students were able to complete the survey on their own time, as the survey was available 24 hours a day, 7 days a week, for a 3-week time period.

Students received an e-mail reminder to complete the surveys every three days until they were completed. After completing the survey each student received a "thank you" page confirming the completion of the survey.

Results

Scale Construction

As with Time 1, Factor Analysis was conducted to determine the underlying latent structure of all the items that assessed *impacts on learning*. Principal axis factoring was conducted using a varimax rotation. Any factor with an eigenvalue over one was retained. In order to be included as part of the factor, items had to load .5 or higher. As with Time 1, the resulting factor structure produced only one factor and was again labeled "teaching effectiveness." Because there was only one factor there was no need for rotation. This factor accounted for 68.105% of variance. All the items loaded positively (See Table 3). The level of internal reliability was once again high as a Cronbach's α =.9521 was achieved and could not be improved by deleting any of the items.

Table 3. Factor Loadings of Survey Items Related to the Impact on Learning for Time #2

Question section (Course or Professor), Question number, and Question focus	**Loading**
Course #1- Structure	.848
Course #2b- Pace	.818
Course #3- Assignment/projects	.807
Course #4- Class discussion	.737
Course #5- Exams	.743
Professor #1- Presentations/explanation	.867
Professor #2b- Enthusiasm	.861
Professor #3b- Stimulate interest	.889
Professor #4- Interactions	.846
Professor #5b- Feedback/comments	.795
Professor #6b- Challenge	.851

Individual Factors

As with Time 1, all "learning" items were once again summed to create a "teaching effectiveness score" and a multiple regression was conducted to assess the independent affect of each "individual factor" (Mauer, et al., 2006). The overall model was significant, F (8, 1676) = 58.838, p = .0001, r^2 = .216. Of the possible Eight individual factors that were thought to possibly influence ratings, seven were significant predictors of teaching effectiveness: these factors included the number of classes missed, hours spent working outside of classroom, how often the student sought the professor's assistance, participated in class discussion, completed assigned readings, and the believed grade for the course. (β,

p, and sr^2 values are presented in Table 4) Of these items "believed grade" accounted for the most unique variance, sr^2 =.216.

Table 4. Time #2- β, p, sr^2 (squared semi-partial correlation) for multiple regression predicting teaching effectiveness score (Covariates: # of missed classes, hours spent per week on course, seeking professors assistance, % of time fully prepared for class, participation in discussion, % of assignment completed, % of course readings completed, and believed grade)

Student Variable	β	**p**	**sr^2**
# of classes missed	.060	.007	.058
Hours spent on course per week	.114	.0001	.103
Seeking professor assistance	.127	.0001	.117
Being fully prepared for class	.164	.0001	.138
Participated in class discussion	.114	.0001	.101
Completed assigned readings	.059	.010	.052
Believed grade	.244	.0001	.216

Establishing External Validity

Effective teaching induces learning; a change in an individual. Three types of changes are possible: a change in knowledge or cognition, a change in skills, and a change in affect or attitude. Higher scores on the teaching evaluation instrument should thus be correlated with indicators of student learning. Three items were included in the survey as learning indicators: did the student believe to know more about the subject after taking the course, did the student's skills improve as a result of taking the course, and did the student's level of awareness about the subject matter increase as a result of taking the course. These three global learning indicators were used in favor of specific measures of learning because of the wide range in content across the courses.

A bivariate correlation at Time 1 revealed significant relationships between the teaching effectiveness score and all three learning indicators; with knowing more r = +.182, p = .001, with skills improved r = +.415, p = .0001, and with awareness increasing r = +.318, p = .0001. Of these three indicators, only skills improved was significantly correlated with grade, r = +.109, p = .026.

Bivariate correlations at Time 2 also revealed significant relationships between the teaching effectiveness score and all three learning indicators; with knowing more r = +.536, p = .0001, with skills improved r = +.597, p = .0001, and with awareness increasing r = +.565, p = .0001. All three of these indicators were significantly correlated with believed grade as well; with knowing more r = +.195, p = .0001, with skills improved r = +.235, p = .0001, and with awareness increasing r = +.220, p = .0001.

These correlations suggest that the teaching effectiveness score is generalizable to other learning related outcomes. Although these three learning measures are significantly correlated to both the teaching effectiveness score and believed grades, in both Time 1 and Time 2 the variance accounted for by the learning indicators was greater for the teaching effectiveness scores than believed grade. While grades are indeed correlated with the teaching effectiveness score, these correlations suggest that the teaching effective score, and not believed grade, is the strongest measure of these learning indicators. These correlations not only provide a measure of external validity, but also add support to Marsh's (1983, 1987) counter to the "grade satisfaction hypothesis".

Discussion

A Standardized Tool

The goal of this pilot study was to design a new student evaluation instrument that would be statistically sound, but also have some practical utility for instructors. The development of the new survey was based in research and focus group feedback. All items were deemed essential to effective teaching (as defined by our working definition) across all disciplines. The instrument was tested both via the traditional pencil and paper format and the more technologically advanced online format. For both tests, factor analysis revealed this survey to measure only one latent factor, termed "teaching effectiveness."

While it is impossible to truly assess teaching effectiveness with just one instrument or assessment, this diagnostic survey appears to have some measure of reliability and validity. The structure of the survey holds together well, as evidenced by the high Cronbach's a in all tests. While measures of convergent validity cannot yet be obtained, feedback from the focus group assessment coupled with the statistical analyses suggest this scale seems to have a high level of face and construct validity. While each item can be assessed individually, the loadings of the latent factors are all very high, supporting the idea that there can be a multidimensional, global assessment of teaching effectiveness that is comprised of a single overarching construct (d'Apollonoa & Abrami, 1997).

Pounder (2007) pointed out that we are at a time in education ripe for exploration into other methods of student evaluation. It might not be the method that should be reconsidered, but rather the construction of the student evaluation instrument that should be evaluated first. McKeachie and Hofer (2001) noted "teaching effectiveness depends not just on what the teacher does, but rather on what the student does" (p.6). Ultimately, what the student does is exert effort to think and learn. If teaching effectiveness is about what the student does, and what the student does is learn, then this new survey has redirected the focus of the evaluation to something *only* students can assess; the impact on their learning. This new focus means that students are now a more reliable source of information.

This focus on learning also appears to reduce the influence of many individual factors and other biases, such as showpersonship. The individual factors that were predictive of the aggregate teaching effectiveness score are all variables where intuitively one would expect to see a relationship. For example, if students feel they know more at the end of a course than before, their skills have improved in the course area, they come to class fully prepared, and they spend a lot of time actively participating in discussions then we would expect there to be a relationship to learning.

We would also expect a relationship to grade with this new survey; the higher the grade the more learning has presumably taken place. Even though numerous items are predictors of teaching effectiveness, it is important to note that the most variance of the teaching effectiveness score that can be explained across both time points is small at best. More importantly, although significant, the unique variance explained for by grade was a very small (sr^2 = .046 at Time 1 and .013 at Time 2) predictor of the teaching effectiveness score. This suggests that, even though grade was a significant predictor, student ratings were not largely driven by the grade they believed they were going to receive in the course, which could reduce the need to "dumb-down" a course (Huemer, 2005) or artificially inflate grades to get high ratings.

A Diagnostic Tool

Kember et al. (2002) have noted that the routine collection of student evaluations provides no guarantee of any improvement in the quality of teaching. Indeed, instructors often find themselves alone in trying to improve their teaching scores. Brookfield (1995) suggests that three primary sources of feedback on teaching exist for any faculty member: the literature on teaching excellence, one's colleagues, and one's students. Excellent examples of using information from student evaluations to improve one's teaching exist in the literature (see, for example, Gallaher, 2000) but can be difficult to find if one is not sure how to look for them. Better institutions create formal mechanisms to empower faculty to improve their teaching; often by offering intervention and professional development programs through "Centers for Teaching Excellence" and the like. Useful information from students can still be a critical tool to aid in self-improvement.

As reported by faculty who took part in the tests, the instrument developed in this article provides a rich set of qualitative and quantitative information about one's classroom teaching efforts. Faculty reported that the focus on learning in the instrument provided true formative feedback for *how and where* to consider improvements. With this new survey, an instructor can have a better understanding of what particular course characteristic is helping the students learn and what is not. An instructor can evaluate the dynamic between the level of a component and its impact on learning (e.g., course pace is fast and it is not helping learning), the relationship between multiple course components, and even between components and student information. Additionally, faculty participants believed this survey would be much more useful in a summative format as an instructor progresses through the tenure and promotion process.

Many students wrote written comments in the open-ended space after each course component, providing a greater amount of written feedback than that of the SET used previously. Faculty found these comments invaluable as students were prompted to comment on particular course characteristics instead of making general comments about the course experience *per se*. Student comments qualified their assessment of course components. Thus, instructors knew better what it was about a particular item that produced a higher or lower evaluation.

Using this instrument in the online format also holds the potential to direct faculty who score below a chosen threshold on any particular item towards helpful learning modules to help them improve. In accessing their SEL evaluations, computer systems could be set to automatically refer low-scoring faculty to directed-learning modules that would assist them in improving specific aspects of their courses. In the long run, this type of assistance could increase the effectiveness of teachers. Universities and colleges make a considerable investment when hiring faculty for tenure-track positions. Integrating professional development and intervention online could decrease the use of SET's as punitive justification for one's dismissal and instead empower faculty to improve on their own.

Though this new survey appears to be valid and meaningful, as McKeachie warns (1997), it should not be used as the only assessment of teaching effectiveness. Future examinations need to explore the utility of the survey with even larger more diverse samples, as well as address how well the survey compares to other forms of teaching effectiveness, such as peer-and self-evaluations, and how it can best be used as *part of* the overall instructor evaluation.

Acknowledgements
We would like to thank Interim Dean Sclafani for commissioning this task force. We would also like to thank all the faulty members who agreed to be part of this study for their openness and commitment to improving education.

References

Angelo, T. A. (1993). "Teacher's Dozen": Fourteen general, research-based principles for improving higher learning in our classrooms. *AAHE Bulletin*, 45, 3-13.

Bain, K. (2004). What Makes Great Teachers Great? *Chronicle of Higher Education,* 50, B7.

Baiocco, S.A., & DeWalter, J.N. (1998) *Successful College Teaching*. New Jersey: Pearson Education.

Basow, S. & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Personnel Evaluation in Education, 18*, 91-106.

Brookfield, S. (1995). *Becoming a Critically Reflective Teacher.* San Francisco: Jossey-Bass.

Cahn, S.M. (1986). *Saints and Scamps: Ethics in Academia, Revised Edition*. New Jersey: Littlefield Adams.

Cashin, W. (1990). Students do rate different academic fields differently. *New Directions for Teaching and Learning, 43,* 113-121.

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

El Hassan, K. (2009). Investigating substantive and consequential validity of student ratings of instruction. *Higher Education Research and Development, 28 (3)*, 319-333.

Entwistle, N. & Tait, H. (1990). Approaches to learning, evaluations of teaching, and preferences for contrasting academic environments. *Higher Education*, 19, 169-194.

Gallaher, T. (2000). Embracing student evaluations of teaching: a case study. *Teaching Sociology*, *28*, 140-147.

Gibbs, G. (1995). *Assessing Student-Centered Courses.* UK: Oxford Brookes University.

Halpern, D. F. (1999). Teaching for critical thinking: Helping college students develop the skills and dispositions of a critical thinker. *New Directions for Teaching and Learning, 80,* 69-74.

Huemer, M. (2005). *Student Evaluations: A Critical Review*. http://home.sprynet.com/~owl1/sef.htm#_14

Kember, D., Leung, D. Y. P. & Kwan, K. P. (2002) Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment and Evaluation in Higher Education, 27*, 411–425.

Lowman, J. (1995) *Mastering the techniques of teaching, 2nd edition.* San Francisco: Jossey-Bass.

Marsh, H. W. (1982). SEEQ: a reliable, valid and useful instrument for collecting students' evaluations of university teaching, *British Journal of Educational Psychology, 52*, 77–95.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics, *Journal of Educational Psychology, 75*, 150–166.

Marsh, H. W. (1987). Students' evaluations of university teaching: research findings, methodological issues, and directions for future research, *International Journal of Educational Research, 11*, 253–388.

Marsh, H. W. (1991). Multidimensional students' evaluations of teaching effectiveness: a test of alternative higher-order structures, *Journal of Educational Psychology, 83*, 285–296.

Marsh, H.W. & Roche, L.A. (1997). Making students' evaluations of teaching effectiveness effective; The critical issues of validity, bias, and utility. *American Psychologist, 52,* 1187-1197.

Mauer, T.W., Beasley, J.J., Long Dilworth, J.E, Hall, A.H., Kropp, J.J., Rouse-Arnett, M., & Taulbee, J.C. (2006). Child and family development students polled: Study examines student course evaluations. *Journal of Family and Consumer Sciences, 98,* 39-48.

McKeachie, W.J. (1997). Student ratings: The validity of use. *American Psychologist, 52,* 1218-1225.

McKeachie, W.J. & Hofer, B.K. (2001). *McKeachie's Teaching Tips: Strategies, Research, and Theory for College and University Teachers, 11th edition*. Houghton Mifflin.

Messick, S. (1989). Validity. In R. L. Lynn (ed.), *Educational Measurement, 3rd ed.* Old Tapan, NJ: Macmillan.

Ory, J.C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? In M. Theal, P. Abrami, & L. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* (New Directions for Institutional research, #109), 27–44. San Francisco: Jossey-Bass.

Perry, W.G. (1970). *Intellectual and Ethical Development in the College Years: A Scheme*. New York: Holt, Rinehart, and Winston.

Pounder, J.S. (2007). Is student evaluation of teaching worthwhile?: An analytical framework for answering the question. *Quality Assurance in Education, 15,* 178-191.

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

Ramsden, P. & Entwistle, N. J. (1981). Effects of academic departments on students' approaches to studying, *British Journal of Educational Psychology,* 51, 368–383.

Richardson, J. T. (2005). Instruments for obtaining student feedback: a review of the literature. *Assessment & Evaluation in Higher Education, 30 (4), 387–415*

Shuman, H. & Presser, S. (1979). The assessment of 'no opinion' in attitude surveys. *Sociological Methodology, 10*, 241 – 275.

Teverow, P. (2006). Another metaphor for teaching excellence: Machiavelli's The Prince. *The Teaching Professor, Jan.,* 3.

Titus, J. (2008). Student ratings in a consumerist academy: leveraging pedagogical control and authority. *Sociological Perspectives*, *51*, 397-422.

Trout, P. (2000). Flunking the Test: The Dismal Record of Student Evaluations. *Academe Online, the Bulletin of the American Association of University Professors, http://aaup.org/AAUP/pubsres/academe/2000/JA/Feat/trou.htm. Accessed 11/10/2010.*UC Regents (2000). *Assessment of Student Learning Gains.* http://mc2.cchem.berkeley.edu/Evaluation/class_ev.html

Voss, R. & Gruber, T. (2006). The desired teaching qualities of lecturers in higher education: a means end analysis. *Quality Assurance in Education, 14,* 217- 242.

Ware, J.E., & Williams, R.J. (1975). The Dr. Fox effect: a study of lecturer effectiveness and ratings of instruction. *Journal of Medical Education, 50,* 149-156.

International Journal for the Scholarship of Teaching and Learning
http://www.georgiasouthern.edu/ijsotl
Vol. 5, No. 1 (January 2011)
ISSN 1931-4744@ Georgia Southern University

Appendix A
Classroom Survey© 2006-2007

Classroom Survev

COURSE CODE:

PROFESSOR'S NAME:

WHY YOU SHOULD COMPLETE THIS EVALUATION
The university is dedicated to continuously improving classroom instruction. As a way of furthering this mission, we value your input regarding your direct experience in this course. **Your responses are part of the overall faculty evaluation process and can help both the university and your professor better understand your classroom experience and the impact it has on your learning.**

INSTRUCTIONS

Please read the instructions at the beginning of each section carefully. Fill in the box that corresponds to your response for each item with either a check mark or an X. Please choose only ONE response for each item, and then write your comments in the spaces provided. **All your responses will be kept anonymous. Completion of this form is voluntary. Faculty will not see your responses until after final grades have been submitted.**

Thank you for completing this survey!

GENERAL INFORMATION
Your participation in the following three questions is optional. The university collects these data with the intention of enhancing all students' learning experiences across majors, sexes, and ethnicities.

1. I am:	**Male** o	**Female** o			
2. My current status at UT is: OTHER: _____	**Freshman** o	**Sophomore** o	**Junior** o	**Senior** o	**Grad Student** o

3. I consider myself to be OTHER: _____	**African-American/ Black** o	**Asian** o	**Caucasian/ White** o	**Hispanic/ Latino** o	**Pacific Islander** o	**Native American** o	**Multi-ethnic** o	**Unknown** o

International Journal for the Scholarship of Teaching and Learning
http://www.georgiasouthern.edu/ijsotl
Vol. 5, No. 1 (January 2011)
ISSN 1931-4744@ Georgia Southern University

THE COURSE – Indicate below how each aspect of the course **impacted your learning** by checking one box for each statement. The response scale for most items ranges from "Did not help my learning" to "Helped my learning a great deal." If **you are unable to evaluate a particular aspect in any way, please choose "Not Applicable."**

	Did not help my learning	Helped my learning a little	Helped my learning adequately	Helped my learning a lot	Helped my learning a great deal	Not Applicable
1. The way this class was structured	❒	❒	❒	❒	❒	❒
COMMENTS:						
2 A. The pace of the course was **D** SLOW **D** MEDIUM **D** FAST						
B. The pace at which this course progressed	**D**	**D**	**D**	**D**	**D**	**D**
COMMENTS:						
3. The class assignments/projects/activities	**D**	**D**	**D**	**D**	**D**	**D**
COMMENTS						
4. A. Class discussions occurred. **D** NEVER **D** RARELY **D** PERIODICALLY **D** FREQUENTLY						
B. The class discussions	**D**	**D**	**D**	**D**	**D**	**D**
MENTS:						
5. The exams	**D**	**D**	**D**	**D**	**D**	**D**
COMMENTS:						

https://doi.org/10.20429/ijsotl.2011.050111

International Journal for the Scholarship of Teaching and Learning
http://www.georgiasouthern.edu/ijsotl
Vol. 5, No. 1 (January 2011)
ISSN 1931-4744@ Georgia Southern University

THE PROFESSOR - Indicate below how each aspect **impacted your learning** by checking one box for each statement. The response scale for most items ranges from "Did not help my learning" to "Helped my learning a great deal." **If you are unable to evaluate a particular aspect in any way, please choose "Not Applicable."**

	Did not help my learning	Helped my learning a little	my learning	Helped my learning a lot	Helped my learning a great deal	Not Applicable
1. The professor's presentations and explanations in class	O	O	O	O	O	O
COMMENTS:						
2 A. The professor seemed to have enthusiasm for the subject. O NO O LOW O MEDIUM O HIGH						
B. The professor's level of enthusiasm for the subject	O	O	O	O	O	O
COMMENTS:						
3. A. The professor stimulated ____ interest in the subject. O NO O LOW O MEDIUM O HIGH						
B. The level at which the professor stimulated interest in the subject	O	O	O	O	O	O
COMMENTS:						
4. The professor's interactions with me	O	O	O	O	O	O
COMMENTS:						
5. A. The professor provided comments and feedback on my work.						

International Journal for the Scholarship of Teaching and Learning
http://www.georgiasouthern.edu/ijsotl
Vol. 5, No. 1 (January 2011)
ISSN 1931-4744@ Georgia Southern University

	Did not help my learning	Helped my learning a little	my learning	Helped my learning a lot	my learning a great deal	Not Applicable
B. The professor's comments and feedback on my work	D	D	D	D	D	❐

COMMENTS:

6. A. This professor_ challenged me to do better.

D NEVER D RARELY
D PERIODICALLy D ENTLY

B. The level at which this professor challenged me	D	D	D	D	D	D

COMMENTS

1. What aspect(s) of your **classroom experience (course, professor, etc.)** helped your learning most?

2. What aspect(s) of your **classroom experience (course, professor, etc.)** could have been changed to help your learning?

International Journal for the Scholarship of Teaching and Learning
http://www.georgiasouthern.edu/ijsotl
Vol. 5, No. 1 (January 2011)
ISSN 1931-4744@ Georgia Southern University

rHE STUDENT – The information in this section is important to your professor for the purposes >f improving teaching. Your responses below will not impact the validity of your responses in :he previous sections. Please answer each statement honestly.

I. Are you either a major/minor in the department in which this course is offered?					YES D	NO D
- Is this a required course?					YES D	NO D
I. believe know more about this ;ubject now than did before took his course.	Strongly Disagree D	Disagree D	Neither Agree or Disagree D	Agree D	Strongly Agree D	
L believe my skills in this area 1ave improved as a result of taking his course.	Strongly Disagree D	Disagree D	Neither Agree or Disagree D	Agree D	Strongly Agree D	
i. believe my awareness of this ;ubject has increased as a result of aking this course.	Strongly Disagree D	Disagree D	Neither Agree or Disagree D	Agree D	Strongly Agree D	
i. How many prior courses have you aken in this department?	None D	1-2 D	3-4 D	5-6 D	7 or more D	
7. How many class meetings did you niss in this course?	None D	1-2 D	3-4 D	5-6 D	7 or more D	

- Approximately how many hours *per week* Jid you spend preparing for this course?	0 hours D	1-3 hours D	4-6 hours D	7-9 hours D	10-12 hours D	13 or more hours D
). How often did you seek the professor's 1ssistance and/or have discussions with 1er/him outside of class?	None D	1-2 times D	3-5 times D	6-9 times D	10-12 times D	13 or more times D

lO. Based on the instructor's xpectations, was fully prepared for % of the class meetings 1ttended.	0°/o D	1-20% D	21-40% D	41-60% D	61-80% D	81-99% D	100% D
II. actively participated in _____%of he class discussions.	0°/o D	1-20% D	21-40% D	41-60% D	61-80% D	81-99% D	100% D
I2. completed _____% of the class lSsignments/projects.	0°/o D	1-20% D	21-40% D	41-60% D	61-80% D	81-99% D	100% D
I3. completed _____% of the course ·eadings.	0°/o D	1-20% D	21-40% D	41-60% D	61-80% D	81-99% D	100% D

I4. believe my final grade in this course will be:	A D	AB D	B D	BC D	C D	CD D	D D	F D