Review on the use of student evaluations of teaching (SETs), with recommendations for assessment of teaching faculty for promotion, tenure, and annual review.

By

The Teaching Effectiveness Committee, Auburn University

Todd D. Steury, School of Forestry and Wildlife Sciences (Former Chair)

Murali Dhanasekaran, School of Pharmacy (Current Chair)

Melissa Blair, College of Liberal Arts

Diane Boyd, Biggio Center

Sierra Bostick, Undergraduate Representative

Katie Boyd, Office of Academic Assessment

Bethany Cleveland, College of Education

Matthew Hall, College of Architecture, Design, and Construction

Katilya Shadrell Harris, School of Nursing

Mandy Harrelson, College of Business

Julie Huff, Strategic Initiative and Communications

David King, College of Science and Mathematics

Roy Knight, College of Engineering

Tom Leathem, College of Architecture, Design, and Construction

Mahmoud Mansour, College of Veterinary Medicine

David Martin, College of Human Science

Lindsey Moseley, Graduate Student Representative

William Pope, School of Nursing

Constance Relihan, Office of Undergraduate Studies

Carolyn Robinson, College of Agriculture

Juliet Rumble, Library

Brandon Simmons, Instructional Technology

Bailey Sullivan, Undergraduate Representative

Emily Wilkins, Graduate Student Representative

Summary

Starting in December of 2017, on the recommendation of its Teaching Effectiveness Committee, Auburn University modified the questions that would comprise the University-required Student Evaluations of Teaching (SETs) questions administered at the end of the semester for the majority of university classes. Part of the impetus for that change was to move away from questions that are intended to be summative evaluations of teaching (i.e., absolute measures of the quality of teaching) in favor of questions that are more formative evaluations of teaching (i.e., specific ways in which teaching can be improved). This change was based on an increasing body of research that demonstrates the inability of SETs to evaluate individual teaching effectiveness in a summative manner. In this report, we briefly review the literature on the utility of SETs. Furthermore, we review the literature for expert advice on how SETs can and should be used and for advice on how teaching effectiveness can be evaluated. We close with specific recommendations for how teachers should be evaluated at Auburn University, via SETs and other measures of teaching effectiveness, for the purposes of promotion, tenure, annual review, and other employment decisions.

Introduction

One of the charges of the University Teaching Effectiveness Committee at Auburn University is to oversee the Student Evaluations of Teaching (SETs) that are conducted at the end of the semester for the majority of undergraduate classes at the university, including the development of the questions that will be used in those SETs (Auburn University, 2016). In 2016, the Teaching Effectiveness Committee began an extended examination of the questions in use at the time (used since 2011) and, finding that revisions were in order, began the process of developing new questions for use in the university-wide SETs. Part of this process included modified focus groups with faculty and students, and was informed by an in-depth review of existing research on SETs, including their validity, reliability, and utility, in order to write questions that would be of broad applicability in their use. In conducting this research, the committee found a substantial body of research suggesting that SETs are generally of limited value for summative (i.e., absolute) evaluations of teaching effectiveness of individual instructors (see below). Thus, the committee set forth to generate questions for the SETs that would be of greatest value in the formative evaluations of teaching; that is, ways in which

teaching effectiveness can be improved over time. The final questions chosen mirrored the 7 principles for good practice in undergraduate education by Chickering and Gamson (1987). In discussions about changes to the SETs, it also became apparent to the committee that many units on campus were using SETs in an inappropriate fashion. Thus, the committee decided to write this report with the goal of providing well-supported information on what SETs should be used for, why use of SET scores for employment decisions such as promotion and tenure is inappropriate in most circumstances, and how administrators can best evaluate instructors in a fair, but efficient manner.

Review of literature on SETs

A review of the existing literature on the validity and utility of SETs suggests that the topic is a highly controversial one, with scientific studies both in support and against the notion that SETs are valid tools for measuring teaching effectiveness (see reviews by Cashin 1995, Benton and Cashin 2012, Berk 2013, Benton and Li 2015). However, a careful examination into the similarities and differences among these studies reveal there are some critical conclusions that can be made about SETs and their utility in the evaluation of individual instructors for employment decisions. Specifically, 1. students are not qualified to evaluate faculty for teaching effectiveness; 2. the evidence for the validity of SETs for measuring individual teaching effectiveness has always been weak and is recently waning; and 3. many sources of bias in SET scores exist.

The literature on SETs uniformly acknowledges that students are not qualified to evaluate the effectiveness of a teacher. Rather SETs are used to gather the collective views of a group of students about their experience in a course taught by a particular faculty member (Arreola 2007, Hativa 2013, Linse 2017). Numerous behaviors and skills define teaching effectiveness - none of which students are qualified to rate - such as a professor's knowledge and content expertise, teaching methods, course design and organization, use of technology, quality of course materials, assessment instruments, and grading practices (Marsh 2007, Ory and Ryan 2001, Svinicki and McKeachie 2011, and Theall and Franklin 2001, Berk 2013).

Second, and as an obvious consequence of the first conclusion, SETs are not good measures of the effectiveness of individual teachers or student learning in individual classes. Notably, there are quite a few studies that, at first glance, seem to suggest that SETs are

positively correlated with teaching effectiveness and student learning; i.e., the validity of SETs (e.g., Wright and Jenkins-Guarnieri 2012; see reviews by Benton and Cashin 2012). However, when examined more carefully, a common theme to these studies is the finding that SETs are useful for measuring teaching effectiveness or student learning *in aggregate* – that is; SETs can be useful when comparing multiple individuals across departments or faculty across programs. However, none of these studies provide strong evidence that SETs can be useful in distinguishing among the teaching effectiveness of individual teachers, especially with scores from a single class. Rather, the majority of research studies on the relationship between SET scores and various measures of student learning have shown low to moderate positive correlations. Indeed, meta-analyses of previous research into the subject found that the correlation coefficient between SET ratings and student learning of course material (as measured from final exam scores in courses with multiple sections, each with their own instructor and receiving their own SET score) ranged from 0.13 to 0.44 (Cohen 1981, 1982, 1983, Feldman 1987, McCallum 1984, Clayson 2009). At first glance, these numbers would suggest that, given the positive correlation, SETs are a valid instrument for measuring student learning. The problem with the relatively low correlations between SET ratings and student learning is that they indicate clearly why SETs are not an effective tool to evaluate individual faculty performance in individual classes. Consider the highest observed correlation of 0.44; this number means that only 20% of the variation in student scores is captured by student learning. The other 80% of the variation in student scores is driven by other factors. The result is that it's entirely possible to have instructors who are very effective at teaching, but get very low scores on SETs, and vice versa (Fig. 1). Thus, results from these studies suggest that SET scores from a single course has little meaning for evaluating the effectiveness of an instructor. At best, SET scores from multiple courses taught over multiple years *may* give *some* indication of the effectiveness of an individual teacher, but due to the limits of the instrument, using them as performance indicators for merit increases, promotion, and tenure is unwise (see below).

Moreover, newer studies on the validity of student ratings has led to decreasing support for the validity of SETs, and increased evidence of bias (reviewed by Nilson 2012, also see Uttl et al. 2017). Indeed, a recent re-analysis of previous meta-analysis studies found that most of them committed the same methodological flaw: they all failed to consider the possibility that the small to moderate SET/learning correlations may be an artifact of small sample sizes in the

majority of studies and small sample bias (Uttl et al. 2017). Specifically, studies with small sample sizes were more likely to have strong correlations than studies with large sample sizes. The reason, as is often noted in meta-analytical studies, is that studies with small sample sizes need strong effects to find significance, and significance is often required to publish results. Consequently, if the true correlation is relatively weak, those results often will not show up in the published literature. Uttl et al. (2017) demonstrated that previous studies were in fact subject to this small-sample size bias, and that adjusting for this bias substantially decreased the estimated correlation coefficient. For example, when Uttl et al. (2017) re-analyzed Cohen's original 1981 data, they found an adjusted (for small sample-size bias) coefficient of just r = 0.27 (originally, it was 0.43 for overall instructor). Uttl et al. (2017) made similar conclusions about Felman's 1989 study (adjusted r = 0.05) and Clayson's 2009 study (adjusted r = 0.06). Uttl et al. (2017) went on to conduct their own meta-analysis, and found no relationship between SET scores and student achievement (adjusted r = -0.02 to -0.04). Other recent meta-analyses of validity data have come to similar conclusions (e.g., see Onweugbuzie et al. 2009).

Third, as indicated by the relatively weak correlation between SET scores and student learning, researchers have identified a number of potential sources of bias in SET scores, including faculty rank (Braskamp and Ory 1994), gender (Centra and Gaubatz 2000), expressiveness (Marsh and Ware 1982), student motivation (Marsh and Dunkin 1992, 1997), whether the course is required or not (Centra 2009), the grade the student expects to receive (Centra 2003), level of the course (i.e., lower- vs. upper-division classes; Braskamp and Ory 1994), class size (Hoyt and Lee 2002a), academic discipline (Hoyt and Lee 2002b), and student workload (Hoyt and Lee 2002a, Centra 2003). Thus, differences in SETs may reflect individual student biases rather than teaching effectiveness *per se*.

The degree to which the prevailing literature supports the notion that SETs are not valid instruments for measuring the effectiveness of individual teachers has important implications for their use. Most notably, if SETs are used for that purpose in making decisions about annual review, merit-based raises, continuation of employment, promotion, or tenure, then departments, colleges, and the university open themselves up to lawsuit over those decisions (Wines and Lau 2006). The problems with the use of SETs in employment decisions is most notable when based on 'global' questions of teaching effectiveness (e.g., "The instructor's overall effectiveness was…"). Berk (2013) reviewed the appropriateness of such 'global' questions for personnel

decisions and found that they failed four critical standards: psychometric standards (such questions tend to have low reliability); representativeness and fairness (it's not fair to evaluate an employee's performance on a single score); professional standards for employee decision (relying on one or two global ratings alone for major summative decisions about faculty performance violates various personnel testing and evaluation standards); and legal standards for employee decisions (use of single ratings along for major summative decisions about faculty performance may violate various laws). Indeed, due to inherent bias that may exist in SET scores, use of such scores for employment decisions may result in discriminatory practices based on age, gender, race, ethnicity, or other protected classes (U.S. EEOC 2010).

Recommendations

Student evaluation scores (especially from a single course) are poor indicators of an instructor's overall effectiveness. For formative evaluations of instructors, ratings should come from two or more courses every term, for at least two years, totaling six to eight courses; if any course has fewer than 10 raters, data from additional classes is recommended (Benton and Cashin 2012). If the evaluations from any one course has a poor participation rate, evaluations from that course should not be used. Indeed, some sources (e.g., Berk 2013) suggest that if scores are to be used for summative employment decisions, a minimum participation rate of 70% is needed. For formative considerations related to course improvement, a lesser 30% participation may be sufficient (Berk 2013). Additionally, global questions about teaching effectiveness should never be used (Berk 2013); instead, the individual scores from all questions is preferred or perhaps a mean or median of the scores from all questions. However, student evaluations are not designed to gather comparative data about faculty (Franklin 2001). Specifically, biases that students may have against instructors of a particular gender, race, or other characteristic are most likely to generate differences in faculty comparisons. Thus, the recommendation of this committee with respect to the use of SETs in evaluations of teaching and student learning are that SET scores should be used largely in a formative manner (i.e., to improve teaching, or at most to evaluate if teaching has improved over time), rather than a summative manner (i.e., as an absolute measure of teaching performance). Additionally, the committee recommends that evaluations of teaching should include several other methodologies, as detailed later in this report, with an emphasis on formative measures, as opposed to

summative ones. The use of SET scores in a summative manner is fraught with limitations, most notably that variation in such scores is often primarily driven by variation in factors unrelated to teaching effectiveness or student learning. At most, an instructor who *consistently* scores poorly on SETs across a variety of courses, taught at a variety of times, and in a variety of different scenarios (e.g., required versus elective courses) *may* have poor teaching effectiveness (Berk 2013, Linse 2017); although such decisions are still unable to overcome potential biases against that instructor because of characteristics of the instructor themselves that are unrelated to teaching (i.e., gender, race, etc.)

However, the committee recognizes the importance of having easily quantifiable and repeatable means of evaluating the performance of teachers. We believe that SETs could still be used in a formative manner for evaluating teaching performance; specifically, regardless of the absolute scores teachers receive on SETs, scores that improve over time are indicative of a teacher that considers student learning seriously, and takes steps to improve their coursework, teaching style, and student outcomes on a regular basis.

We caution, however, that even when used in this formative manner SETs should not be used exclusively for any purpose related to employment. For both improvement purposes and for personnel decisions, the use of multiple methods of teaching evaluation involving multiple sources of data should be used (Arreola 2007, Berk 2006, 2009, 2013, Benton and Cashin 2012, Braskamp and Ory 1994, Centra 1993). Such methods may include peer review of course material, peer review of course instruction (an example of a peer-review evaluation rubric is attached; appendix 1), review of metrics of teaching effort by the department chair or other supervisor (e.g., are course policies and procedures followed; are preparations made and grades submitted in a timely manner, etc.), review by expert outside sources such as the Biggio Center, exit and alumni ratings, employer ratings, teaching scholarship, teaching awards, learning outcome measures, teaching portfolios, or even self-evaluation of teaching (Berk 2006, 2013). Rubrics could be generated for any or all of these metrics that provide easily-used scores for quantifiable evaluation of teaching performance. Importantly, to ensure that the evaluation system adopted is credible and acceptable, faculty members must have a strong hand in its development.

Literature Cited

Arreola, R.A. 2007. Developing a comprehensive faculty evaluation system: a guide to designing, building, and operating large-scale faculty evaluation systems. Anker Publishing, Boston, MA.

Auburn University. 2016. *Auburn University Faculty Handbook*, sites.auburn.edu/admin/universitypolicies/Policies/AuburnUniversityFacultyHandbookPolicies.ppdf. Accessed October 9, 2018.

Basow, S.A., and J.L. Martin. 2012. Bias in student evaluations. *In* M.E. Kite, editor. Effective evaluation of teaching: a guide for faculty and administrators. E-book retrieved on October 26th, 2018 from The Society for the Teaching of Psychology website, http://teachpsych.org/ebooks/evals2012/index.php

Benton, S.L, and W.E. Cashin. 2012. Student ratings of teaching: a summary of research and literature (IDEA Paper No. 50). Manhattan, KS: The IDEA Center at Kansas State University, Center for Faculty Evaluation and Development. Accessed on October 24th, 2018, from https://www.ideaedu.org/Research/IDEA-Paper-Series

Benton, S.L., and D. Li. 2015. Validity and reliability of IDEA *Teaching Essentials* (IDEA Research Report #8). Manhattan, KS: The IDEA Center at Kansas State University, Center for Faculty Evaluation and Development. Access on October 29th, 2018, from https://www.ideaedu.org/Research/Research-Technical-Reports

Berk, R.A. 2006. Thirteen strategies to measure college teaching: A consumer's guide to rating scale construction, assessment, and ecision making for faculty, administrators, and clinicians. Stylus Publishing, Sterling, Virginia

Berk, R.A. 2009. Using the 360-degree multisource feedback model to evaluate teaching and professionalism. Medical Teacher 31:1073-1080.

Berk, R.A. 2013. Top 10 flashpoints in students ratings and the evaluation of teaching: what faculty administrators must know to protect themselves in employment decisions. Sterling, Virginia: Stylus.

Boring, A., Ottoboni, K., & Stark, P.B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Braskamp, L.A., and J.C. Ory. 1994. Assessing faculty work: Enhancing individual and institutional performance. Jossey-Bass, San Francisco.

Cashin, W.E. 1995. Student ratings of teaching: The research revisited (IDEA Paper No. 32). Manhattan, KS: The IDEA Center at Kansas State University, Center for Faculty Evaluation and Development.

Centra, J.A., and N.B. Gaubatz. 2000. Is there gender bias in student evaluations of teaching? The Journal of Higher Education 71:17-33.

Centra, J.A. 2003. Will teachers receive higher student evaluations by giving higher grades and less course work? Research in Higher Education 44:495-518.

Centra, J.A. 2009. Differences in responses to the Student Instructional Report: Is it bias? Educational Testing Service, Princeton, NJ.

Clayson, D.E. 2009. Student evaluations of teaching: are they related to what students learn? A meta-analysis and review of the literature. Journal of Marketing Education 31:16-30.

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

Cohen, P.A. 1982. Validity of student ratings in psychology courses: a research and synthesis. Teaching of Psychology 9:78-82.

Cohen, P.A. 1983. Comment on A selective review of the validity of student ratings of teaching. The Journal of Higher Education 54:448-458.

Chickering, A.W., and Z.F. Gamson. 1987. Seven principles for good practice in undergraduate education. American Association for Higher Education Bulletin 3:3-7.

Feldman, K.A., 1989. The association between student ratings of specific instructional dimensions and student achievement: refining and extending the synthesis of data from mutisection validity studies. Research in Higher Education 30:583-645.

Franklin, J. 2001. Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. New Directions for Teaching and Learning 87:85-100.

Hativa, N. 2013. Student ratings of instruction: a practical approach to designing, operating, and reporting. Oron Publications.

Hoyt, D.P., and E. Lee. 2002a. Basic data for the revised IDEA system (IDEA Paper No. 12). Manhattan, KS: The IDEA Center at Kansas State University, Center for Faculty Evaluation and Development.

Hoyt, D.P., and E. Lee. 2002b. Disciplinary differences in student ratings (IDEA Paper No. 13). Manhattan, KS: The IDEA Center at Kansas State University, Center for Faculty Evaluation and Development.

Linse, A.R. 2017. Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. Studies in Educational Evaluation 54:94-106.

Marsh, H.W. 2007. Student's evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R.P. Perry and J.C. Smart, editors. The scholarship of teaching and learning in higher education: an evidence-based perspective. Springer, Dordrecht, The Netherlands.

Marsh, H.W., and M.J. Duncan. 1992. Students' evaluations of university teaching: A multidimensional perspective. In J.C. Smart, editor. Higher education: Handbook of theory and research. Volume 8. Agathon Press, New York.

Marsh, H.W., and M.J. Duncan. 1997. Students' evaluations of university teaching: A multidimensional perspective. In R.P. Perry and J.C. Smart, editors. Effective teaching in higher education: research and practice. Agathon Press, New York.

Marsh, H.W., and J.E. Ware. 1982. Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox effect. Journal of Educational Psychology 74:126-134.

Matthew, P. A. (2016). Wr*itten/Unwritten: Diversity and the hidden truths of tenure*. Chapel Hill, NC: University of North Carolina Press.

McCallum, L.W. 1984. A meta-analysis of course evaluation data and its use in the tenure decision. Research in Higher Education 21:150-158.

Nilson L.B. 2012. Time to raise questions about student ratings. *In* J.E. Groccia and L. Cruz, editors. To improve the academy: Resources for faculty, instructional, and organizational development. Jossey-Bass, San Francisco.

Onwuegbuzie, A.J., L.G. Aniel, and K.M.T. Collins. 2009. A meta-validation model for assessing the score-validity of student teaching evaluations. Quality & Quantity 43:197-209.

Ory, J.C. and K. Ryan. 2001. How do student ratings measure up to a new validity framework? *In* M. Theall, P.C. Abrami, and L.A. Mets, editors. The student ratings debate: Are they valid? How can we best use them? Jossey-Bass, San Francisco.

Svinicki, M. and W.J. McKeachie. 2011. McKeachie's teaching tips: Strategies, research, and theory for college and University teachers. Wadsworth, Belmont, CA.

Theall, M., and J.L. Franklin. 2001. Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *In* M. Theall, P.C. Abrami, and L.A. Mets, editors. The student ratings debate: are they valid? How can we best use them? Jossey-Bass, San Francisco.

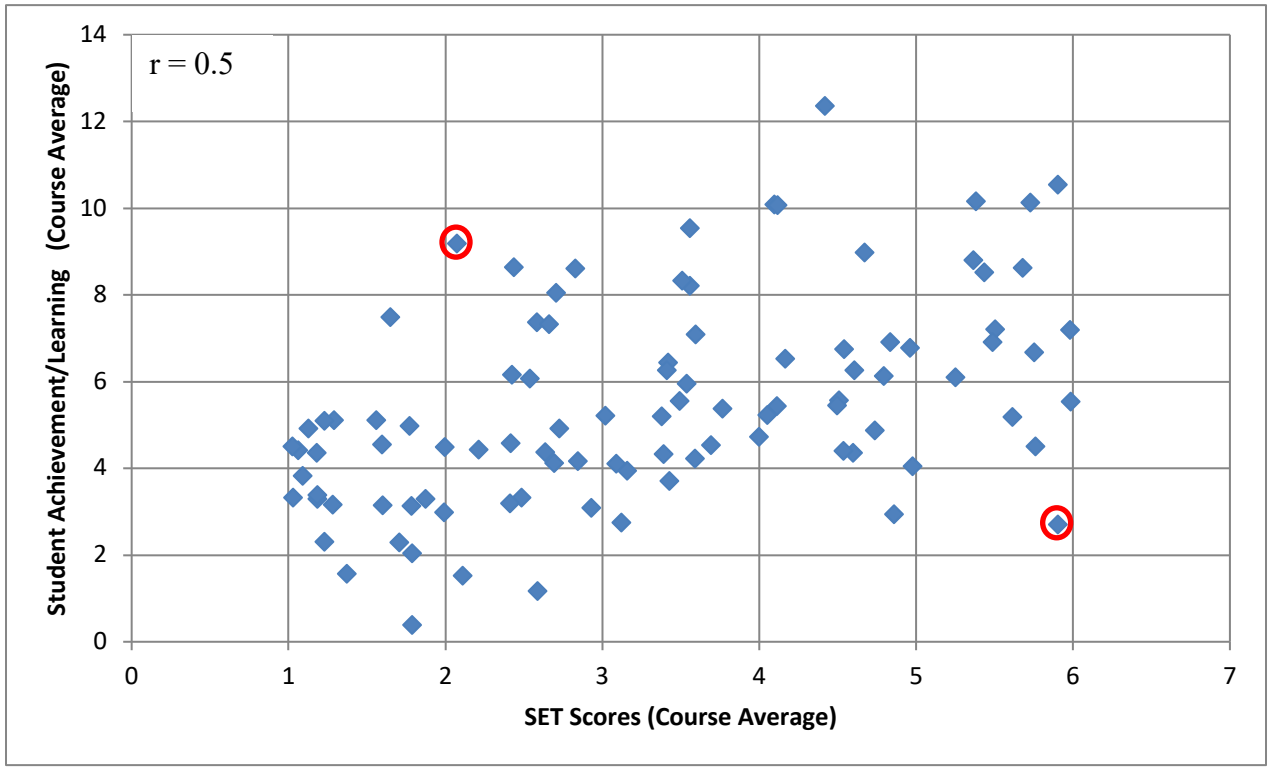
Uttl, B., C.A. White, D.W. Gonzalez. 2017. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. Studies in Educational Evaluation 54:22-42.

U.S. Equal Employment Opportunity Commission (EEOC). 2010. Employment tests and selection procedures. Accessed on October 29, 2018, from http://www.eeoc.gov/policy/docs/factemployment_procedures.html

Wright, S.L., and M.A. Jenkins-Guarnieri. 2012. Student evaluations of teaching: combining the meta-analyses and demonstrating further evidence for effective use. Assessment & Evaluation in Higher Education 37:683-699.

Wines, W.A., and T.J. Lau. 2006. Observations on the folly of using student evaluations of college teaching for faculty evaluation, pay, and retention decisions and its implications for academic freedom. William & Mary Journal of Women and the Law 13:167-202.

Figure 1. This figure represents a hypothetical scatter plot of data with a 0.5 correlation (a bit better than the highest correlation found in studies of well-validated student evaluation instruments). The data clearly depict a positive relationship between evaluation scores and student achievement. Yet, if we pick any two individual courses, the results could be flipped. In the case of the two data points highlighted in red, the instructor with the lower evaluation score actually generated substantially greater student learning than the instructor with the higher score.



Appendix 1

Example Peer Review Evaluation Sheet[4]

Scale:

1 =Very Poor; needs serious substantial improvement
2 =Poor; needs much improvement
3= Good; needs; needs a fair amount of improvement
4 ''' Very good; needs a little improvement
5 = Excellent; needs no improvement

Content and Delivery	1	2	3	4	5	N/A	Comments
Appropriate use of time (begins and ends on time)							
Provides introduction/overview of topic/daily goals							
Appropriate level of presentation (Depth and breadth)							
Claritv of presentation (Seems prepared; explains jargon)							
Relevance of information (Stays on topic)							
Knowledge (Uses citations; answers questions clearly)							
Logical flow (Well organized, useful transitions)							
Pace of presentation							
Poses appropriate and clear questions							
Repeats Students' Questions and Comments							
Use of relevant examples in presenting topic							
PowerPoint (Avoids direct reading off of screen)							
PowerPoint (Grammar and spelling)							
PowerPoint (Clarity-Proper font size and visual clarity)							
Use of Demonstration/Links to Concepts							
Use of Active Learning Techniques						o	
Handouts (Useful in understanding topic)							
Provides conclusion/take home message							
Physical Presence	1	2	3	4	5	N/A	Comments
Makes eye contact with general audience							
Makes eye contact while speaking to individuals							
Facial expression							
Movement about room							
Posture							
Professional attire							
Use of appropriate hand gestures							
Voice-Audible							
Voice-Variation in inflection and tone							
Voice-Appropriate pace of speaking							
Social Presence	1	2	3	4	5	N/A	Comments
Composure/Confidence							
Reinforces student participation							
Relaxed teaching style (may include sense of humor)							
Engaging (Interesting and informative)							
Respectful							
Use of student names							

Other Comments:

_____
_____
_____
_____
_____

[4] Developed by William Buskist, Auburn University Department of Psychology. Used by permission of author.