Reliability and validity of an evaluation tool for the online class

Prof. Amany Saleh, Ph. D. [asaleh@astate.edu]
Arkansas State University [http://www.astate.edu]
Educational Leadership, Curriculum, and Special Education
P.O. Box 2781, State University, AR

Assist. Prof. Marcia Lamkin, Ed. D. [m.lamkin@unf.edu]
University of North Florida [http://www.unf.edu]
Department of Leadership, Counseling & Instructional Technology,
1 UNF Drive, Jacksonville, FL 32224

Abstract

As institutions of higher education experience a dramatic rise in the demands for online classes, faculty members are at a loss for available tools effectively to evaluate their teaching practices. The authors of this article developed an instrument to give higher education faculty reliable feedback on their online classes. The authors developed an instrument that is unique to the online classroom and addresses issues that evaluation tools for traditional classes cannot address, such as course delivery, instructor's online input, and efficiency of the medium. In this article, the authors report on the reliability and validity of this instrument.

Key Words

Online Teaching, Course Evaluation, Evaluation of Online Teachers

List of Topics

- Online teaching
- Online course evaluation
- Designing a summative evaluation instrument for the online course
- Reliability and validity of the online course evaluation instrument

Introduction

Enrolment in online courses has drastically increased in the last decade. This increase has led to the intensified need for course evaluation tools that are developed specifically for online courses. Over the last few years, many instructors have expressed their dissatisfaction with the inadequacy of traditional course evaluations to provide them with useful feedback to improve their teaching methods in their online classes. The authors of this article developed a course evaluation instrument designed to address the needs of online educators.

McVay Lynch (2002) contended that one of the most difficult obstacles to overcome in the use of students' surveys to evaluate an online course was the students' inability to separate among the course content (materials, assignments, and activities), the instructor's style and personality, and the technical course delivery methods. She stated, "A sticky subject at most schools is the evaluation of the instructor. In the university system, end-of-course student evaluations often serve for promotion and tenure purposes. Consequently, the creation, validating, and reliability of any instruments used for this purpose is of high concern to faculty" (p. 134).

Palloff and Pratt (2003) criticized the use of evaluation tools from traditional face-to-face classes in online classes since they fail to assess the instructors' ability to build learning communities for independent and autonomous learners. They argued that online class evaluation tools should assess faculty members' abilities to engage students in the course, to give meaningful feedback to their students, and to be responsive to students' needs. The authors of this paper developed this online course evaluation tool with these concerns in mind.

Literature Review

Course Evaluation

The online class focuses on building learning communities and facilitating learners' autonomy and independence, which course evaluation tools must address. Palloff and Pratt (2003) argued that online course evaluations should measure instructors' engagement in the course, quality of feedback, responsiveness to questions, support and assistance with projects, and assignments. They also maintained that summative evaluation should be used in the online class but not as the only measure of the effectiveness of the course.

Koontz, Li, and Compora (2006) defined evaluation as the "process of defining, obtaining, and providing useful information to make informed decisions that will enhance the teaching/learning process" (p. 147). They criticized the summative evaluation as it is practiced in higher education because that evaluation fails to provide useful information to online instructors to make informed decisions. Koontz et al. (2006) contended that most instruments ask students to respond to general statements which elicit no specific comments. They recommended that online summative evaluation tools should be designed specifically to measure the effectiveness of the instruction; the efficiency or the time required to learn the materials; the objectives of the coursework; and the attitude of the students toward course content, instruction, and course requirements.

Cooper (2000) pointed out the importance of online course evaluation when she stated that, "Student evaluations help determine the effectiveness of the various components of an online course and address areas that may need revision. They also communicate to students that their input is valuable" (p. 89). Similarly, Lorenzetti (2006) argued that the current course evaluation tools used by higher education institutes are very broad in scope and fail to give instructors feedback that can be used to improve their course delivery.

McKeachie and Svinicki (2006) maintained that online course assessments should provide feedback to instructors on ways that learning "can be facilitated." The assessment, McKeachie & Svinicki (2006) contended, should inform the teacher "how well the students are meeting the objectives." Cooper (2000), Hoffman, (2003), and Lorenzetti (2006) all criticized the use of traditional courses' evaluation tools in online courses. They agreed that there is a great need for course evaluations that are specifically designed for online courses.

Hoffman (2003) agreed that online course evaluation has been receiving increased attention from institutions of higher education over the last few years. In his study, Hoffman asked such institutions to report their use of online course evaluation tools: he found an increase of eight percent among higher education institutes' use over the span of one year. However, he contended that the large majority of such institutions still rely on paper and pencil course evaluation instruments for all classes, both traditional and online.

Palloff and Pratt (2003) listed a number of elements that should be included in a summative evaluation tool for online coursework. They argued that these evaluation items should focus not only on the instructor's performance but on the total experience of the online learner in the course. These elements are:

- The overall online course experience;
- Orientation to the course and course materials;
- The content, including quantity of materials presented and quality of presentation;
- Discussion with other students and the instructor;
- Self-assessment of level of participation and performance in the course;
- The courseware in use, ease of use, and ability to support learning in this course;
- Technical support; and
- Access to resources (p. 98-99).

The authors of this paper recognized the need to develop a course evaluation tool that was different from those which have been used in traditional courses. This instrument took into consideration the fact that the nature of communication among class participants in the online class was different from that in the face-to-face class. Palloff and Pratt (1999) wrote,

In the online classroom, however, the instructor is represented predominantly by the text. Just as with their students, an instructor's engagement with the material and the course is demonstrated through the number, length, and quality of his or her posts. In many cases, the students and instructor may never meet. The physical manifestation of the instructor may be a photograph on a homepage. Although this creates a difficult evaluation process, it also serves, on some level, to make the feedback received from students more valuable, as it relates directly to their experience of the course and the materials they have studied rather than reflecting the personality of the instructor (p.153).

The authors of this paper developed an instrument that includes consideration for the nature of communication among class participants in the online class. This new tool provides feedback on the efficacy of the instructor and the utility of the course from the students' point of view. The instrument elicits students' feedback with regard to four areas: the course delivery methods; materials and instruction; communication among instructor, students, and peers; and support provided for students during the course.

Reliability

Much of the research to establish reliability for newly constructed instruments has been done in the fields of medicine and psychology. A large number of these projects focused on survey instruments designed to measure quality of life under specific circumstances. Rich, Nau, and Grainger-Rousseau (1999) modified an existing questionnaire more quickly to measure quality of life with asthma. Bradley and colleagues (1999) designed an instrument to measure the impact of diabetes on quality of life. Damiano and others (2000) designed and tested a similar instrument to measure patient quality of life with Parkinson's disease. Coyne and others (2002) designed and tested still another questionnaire designed to measure quality of life with overactive bladder symptoms.

Other researchers have worked recently to establish reliability and validity for new instruments in the realm of health and mental health. Bethell, Peck, and Schor (2001) designed a survey to assess health care provisions for well-child care. Seymour and colleagues (2001) tested the validity of an existing questionnaire to measure health issues among older patients with cognitive impairments. Quintana and colleagues (2003) translated and tested the reliability of a Spanish version of the Hospital Anxiety and Depression Scale, an established instrument in its English version. Obayashi, Bianchi, and Song (2003) measured the reliability and validity of nutrition knowledge, socio-psychological factors, and food label use scales from an earlier diet and health knowledge survey. Finally, McMillan, Bradley, Gibney, Russell-Jones, and Sönksen (2003) evaluated two health status measures in adults with growth hormone deficiencies. Most frequently, these researchers all employed Cronbach's alpha as the primary measure of reliability, with a minimum acceptable alpha coefficient value of 0.70.

More closely aligned to the work in question were the recent efforts to construct surveys designed to measure perceptions or attitudes. Walker, Phillips, and Richardson (1993) surveyed a Native American population about minority recruitment to programs of teacher education and employed Cronbach's α to determine internal consistency of the survey instrument. Dowson and McInerney (1997) designed and tested a new instrument to measure in Australian educational settings students' achievement goals and learning strategies. These researchers used both Cronbach's α and factor analysis to establish reliability in their instrument. Kvaerner, Moen, Haugeto, and Mair (2000) studied parental satisfaction after pediatric outpatient surgery, employing both Cronbach's α and Pearson's correlation coefficients to show reliability in their instrument. Meredith, Wenger, Harada, and Kahn (2000) developed a shortened scale, based on a long and unwieldy established instrument, to measure acculturation among Japanese Americans. Cronbach's α was used, but these researchers relied more heavily on factor analysis to demonstrate the reliability of the shortened instrument. Through the use of Cronbach's α, correlation coefficients, and unrotated factor loadings, McGuiness and Sibthorpe (2003) tested a measure of the coordination of health care services. Coyle, Saunderson, and Freeman (2004) designed and evaluated a questionnaire to measure differing attitudes about learning disabilities, piloting the questionnaire among dental and social policy

graduate students and using Cronbach's α across both total results and dental and social policy subgroups.

Methodology

Development of the evaluation

The impetus to create an instrument designed specifically for students to evaluate online classes was occasioned by two desires: the desire better to understand student satisfaction or frustration with the requirements of online coursework and the desire to document online teaching in a way similar to the way that universities document traditional face-to-face teaching. In order to draft the initial evaluation form, these authors examined a number of existing course evaluation forms, drew from past feedback during less formal exchanges with online students over the past seven years, and solicited the input of colleagues who also taught online classes. Potential evaluation questions were narrowed to thirty total items which fell into four categories: course webpage, course structure and content, course instructor, and overall course evaluation; plus one "global" coordination item that summarized students' reaction to the entire course: "The course met my educational needs."

Pilot administration

In order to pilot the original instrument during the spring and summer of 2006, the pilot evaluation form was distributed electronically to seventy-eight students who had participated in four classes during the spring and summer semesters of 2006 at a large public university in the mid-south. The survey was made available through a commercial online service which guaranteed anonymity to participants but provided full details to the researchers on each completed survey. All responses to this pilot course evaluation were maintained confidentially, as would be the responses to traditional course evaluations. The students to whom the pilot evaluation was distributed were predominantly white female graduate students who were pursuing a Masters' degree in curriculum studies. Of the 78 students invited to participate in this pilot study, 58 (74%) responded and completed the evaluation form in full. Response data were entered in the Statistical Package for the Social Sciences, Version 14.0, one variable per item on the pilot evaluation form, plus one item with reverse coding for the final item on the pilot evaluation form. The final item was originally worded so that the "sense" of the answers was in the opposite order as the sense of the other twenty-nine items: testing was completed first with the original coding and then with the reverse coding.

In order to provide assurance that there were no disparities between the two semesters of survey administration or between courses in either of the semesters, t-tests and simple analysis of variance tests were run among all combinations of those participants. No statistically significant differences were discovered among participants by class groups or by semesters.

Results

Validity

Two of the most important and frequently used categories of validity are content validity and construct validity. Content validity reveals whether an instrument truly reflects the "universe" of items in the subject that the instrument claims to measure; while construct validity demonstrates that the instrument measures a definable underlying psychological construct. Although researchers need only to establish one type of validity for a given instrument, these researchers established both content and construct validity for this new evaluation form: both professors and students who have worked online were consulted in order to determine whether this evaluation form asked and provided opportunity to answer the most pertinent questions about online coursework, and student responses on the pilot administration of the evaluation were examined in comparison with other feedback that the students provided to the professors in order to determine whether the evaluation form actually measured the construct of student satisfaction with online coursework. In both cases, the pilot evaluation stood the tests: this instrument demonstrated both content and construct validity.

Reliability

Statistical analyses to measure reliability have been long established. Through the use of these statistical tests, researchers can determine the extent to which the items in an instrument are related to one another, the level at which all items relate to a global "coordination" item on the pilot evaluation instrument ("The course met my educational needs."), an overall idea of internal consistency (repeatability) of the scale as a whole, and specific problem items that need to be reworded or excluded from the instrument in future administrations. For these operations, these researchers used a full set of Spearman's *rho* correlation coefficients, Cronbach's alpha coefficient of internal consistency, and Cronbach's α coefficient when each item was deleted from the total scale. Spearman's *rho* was applied because, in this pilot administration, the minimum ratio of cases to variables (10.4 to 1) could not be met: Spearman's *rho* better evaluates the relationships among responses from small samples of respondents.

Strong Spearman's *rho* correlation coefficients among the items in each of the four subsets on the evaluation instrument plus strong correlation between each item and the "coordination" item ("course met educational needs") were desired. Within each of the four response subsets, each item in the subset correlated significantly to each of the other items with only three exceptions. In the Course Web Pages subset, neither "The web links were relevant." nor ""I was able to interact effectively with the instructor." correlated with "I was able easily to access the course information at the beginning of the course." In the Overall Course Evaluation, the final question on the pilot evaluation, "I prefer to have face-to-face classes." did not correlate with "The course met my educational needs."

With the exception of the two relationships that failed to correlate in the Course Web Pages subset, correlation coefficients ranged from .365 to .764, with 11 of the 13 remaining correlations exceeding .40.In the Course Content and Structure subset, all the Spearman's *rho* values held statistical significance, and the correlation coefficients ranged from .260 to .935, with 33 of the 36 significant correlations exceeding .40.In the Instructor subset, all the Spearman's *rho* values held statistical significance, and the correlation coefficients ranged from .336 to .875, with 65 of the 67 total correlations exceeding .40.With the exception of the one relationship that failed to correlate in the Overall Course Evaluation subset, the two remaining correlation coefficients equaled .433 and .435.

All but one evaluation item was statistically significantly correlated to the global coordination item on the pilot instrument. Responses to "The course met my educational needs." did not correlate significantly to the final item, "I prefer to have face-to-face classes." (p = .466). Spearman's *rho* correlation coefficients between the other evaluation items and that coordination item exceeded .40 in twenty-seven of the remaining twenty-eight items (range .365 - .882).

Cronbach's alpha for the total thirty items was .956 (high internal consistency) with items coded as marked, .964 (high internal consistency) with the final item coded in reverse to align with the scoring sense of the other twenty-nine items. The Cronbach's alpha formula determines the extent to which all items on an instrument measure the same underlying notion, or the extent to which all items on the instrument are internally consistent. In this case, the researchers wanted all items on the evaluation to measure satisfaction with specific components of the online course. The alpha formula is based on repeated comparisons between the scores of individual items and the overall score: the more similar these scores are, the more accurately each item actually measures one part of the overall notion of satisfaction with the course. The maximum possible value for Cronbach's alpha is 1.0, which would indicate a "perfect" correlation between the scores of all the individual items and that one notion of satisfaction, so the value here of .956 or .964 indicates a very strong correlation.

Cronbach's alphas were then recalculated with each single item removed in turn. This procedure allowed the researchers to determine whether any single items had powerfully influenced the original calculation. The alpha values of each recalculation should remain close to the original result. The resulting alpha correlations for all tests remained high, each exceeding .953. With original coding maintained on the final item, the range of Cronbach's alpha was .953 to .966 (all high internal consistency) with one item removed from each statistical test.

Conclusion

An important factor in developing evaluation surveys is to reach a consensus among instructors on the factors that constitute good teaching in the online classroom. Instructors must be clear on the expectations for communication between them and the students, on time limitations, and on the nature of assignments that can be accomplished in such a class.

The authors of this instrument provide a statistically valid tool for online educators which gives them reliable feedback on their teaching as perceived by their students. Based on the increased need for such tools in online classes, such an instrument can be a valuable tool for institutions of higher education (Hoffman, 2003; McVay Lynch, 2002; Lorenzetti, 2006).

The failure of very few evaluation items to correlate in the process of this pilot application could be due to the fact that this limited group of students perceived the items to ask unrelated questions. Participating students might have perceived that neither "The web links were relevant." nor "I was able to interact effectively with the instructor." related directly to their experience in the opening couple weeks of the online course ("I was able easily to access the course information at the beginning of the course."), and that disconnect might explain the lack of correlations among these survey items. This discrepancy could also be attributed to the fact that students were not required to read the links to be successful in the course, but rather to access them as an additional resource. The fact that many of the survey participants were first-time online students might explain their perception of e-mail and discussion boards as ineffective tools of communication as compared to face-to face communication with the instructor. These survey items in particular must be monitored in future applications of the instrument.

The item "I prefer to have face-to-face classes." also did not correlate with the global "coordination" item, "The course met my educational needs." The preference item was the only item on the survey worded originally to code in the opposite direction as the other twenty-nine items: the authors tested this item both as it was written and with reversed coding. For many of these students, the courses at hand were their first online course experiences: their responses to "I prefer to have face-to-face classes." may have been affected by the newness of the experiences. Alternately, students may have perceived their responses to "I prefer to have face-to-face classes." to be comments about the instructor or the process of the course rather than an overall comment about the online experience, and their bias might have changed their responses to this item. This survey item, like the two others that failed to correlate, must be monitored in future applications of the instrument. If these items continue to fail to correlate, then they should be reworded or eliminated from the survey instrument.

These two researchers both continue to use this pilot instrument in their online courses and have begun to recruit other instructors to use the instrument as well. Additional input from students who participate in online classes will serve to clarify the reliability of evaluation items for the purpose of summative evaluation in the context of online instruction.

References

[1] Angelo, T., & Cross, K. P. (1993). *Classroom assessment techniques*. San Francisco, CA: Jossey-Bass Publishers.

[2] Bethell, C., Peck, C., & Schor, E. (2001, May). Assessing health system provision of well-child care: The Promoting Healthy Development Survey. *Pediatrics*, 107 (5), 1084-1094.

[3] Bradley, C., Todd, C., Gorton, T., Symonds, E., Martin, A., Plowright, R. (1999). The development of an individualized questionnaire measure of perceived impact of diabetes on quality of life: The ADDQoL. *Quality of Life Research*, 8, 79-91.

[4] Brookfield, S.D. (1995). *Becoming a critically reflective teacher*. San Francisco, CA: Jossey-Bass Publishers.

[5] Cooper, L. (2000). Online courses: Tips for making them work. *THE Journal*, 27 (8), 86-92.

[6] Coyle, C., Saunderson, W., & Freeman, R. (2004). Dental students, social policy students and learning disability: Do differing attitudes exist? *European Journal of Dental Education*, 8, 133-139.

[7] Coyne, K., Revicki, D., Hunt, T., Corey, R., Stewart, W., Bentkover, J., Kurth, H., & Abrams, P. (2002). Psychometric validation of an overactive bladder symptom and health-related quality of life questionnaire: The OAB-q. *Quality of Life Research*, 11, 563-574.

[8] Damiano, M., McGrath, M. M., Willian, M. K., Snyder, C. F., LeWitt, P.A., Reyes, P. F., Richter, R. R., & Means, E. D. (2000). Evaluation of a measurement strategy for Parkinson's disease: Assessing patient health-related quality of life. *Quality of Life Research*, 9, 87-100.

[9] Dowson, M., & McInerney, D. (1997, March 24-28). The development of goal orientation and learning strategies survey (GOALS-S): A quantitative instrument designed to measure students' achievement goals and learning strategies in Australian educational settings. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

[10] Harasim, L., Hiltz, S. R., Teles, L., & Turoff, M. (1996). *Learning networks*. Cambridge, MA: MIT Press.

[11] Hoffman, K. M. (2003). Online course evaluation and reporting in higher education. *New Directions for Teaching and Learning*, 96 (3), 25-29.

[12] Koontz, F. R., Li, H., & Compora, D. P. (2006). Designing effective online instruction: A handbook for web-based courses. Oxford, UK: Rowman & Littlefield Education.

[13] Kvaerner, K. J., Moen, M. C., Haugeto, O., & Mair, I. W. S. (2000). Pediatric otolaryngology – parental satisfaction study in outpatient surgery. *Acta Otolaryngol Supplement*, 543, 201-205.

[14] Lorenzetti, J. P. (2006, March 15). Course evaluation project is model for content assessment (Distance Education Report). Magna Publications Inc.

[15] McKeachie, W. J., & Svinicki, M. (2006). McKeachie's teaching tips: Strategies, research, and theory for college and university teachers. Boston, MA: Houghton Mifflin Company.

[16] McGuiness, C., & Sibthorpe, B. (2003). Development and initial validation of a measure of coordination of health care. *International Journal for Quality of Health Care*, 15 (4), 309-318.

[17] McMillan, C. V., Bradley, C., Gibney, J., Russell-Jones, D. L., Sönksen, P. H. (2003). Evaluation of two health status measures in adults with growth hormone deficiency. *Clinical Endocrinology*, 58, 436-445.

[18] McVay Lynch, M. (2002). *The online educator: A guide to creating the virtual classroom*. London: Routledge-Falmer.

[19] Meredith, L. S., Wenger, N., Harada, N., & Kahn. K. (2000). Development of a brief scale to measure acculturation among Japanese Americans. *Journal of Community Psychology*, 28 (1), 103-113.

[20] Obayashi, S., Bianchi, L. J., & Song, W. Reliability and validity of nutrition knowledge, socio-psychological factors, and food label use scales from the 1995 Diet and Health Knowledge Survey. *Journal of Nutrition Education and Behavior*, 35 (2), 83-92.

[21] Palloff, R. M., & Pratt, K. (1999). *Building learning communities in cyberspace: Effective strategies for the online classroom*. San Francisco, CA: Jossey-Bass Publishers.

[22] Palloff, R.M., & Pratt, K. (2003). *The virtual student: A profile and guide to working with online learners*. San Francisco, CA: Jossey-Bass Publishers.

[23] Quintana, J. M., Padierna, A., Esteban, C., Arostegui, I., Bilboa, A., & Ruiz, I. Evaluation of the psychometric characteristics of the Spanish version of the Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*, 107, 216-221.

[24] Rich, L. D., Nau, D. P., Grainger-Rousseau, T. J. (1999). Evaluation of patients' health-related quality of life using a modified and shortened version of the Living With Asthma Questionnaire (ms-LWAQ) and the Medical Outcomes Study, Short-Form 36 (SF-36). *Quality of Life Research*, 8, 491-499.

[25] Seymour, D. G., Ball, A. E., Russell, E. M., Primrose, W. R., Garratt, A. M., & Crawford, J. R. (2001). Problems in using health survey questionnaires in older patients with physical disabilities: The reliability and validity of the SF-36 and the effect of cognitive impairment. *Journal of Evaluation in Clinical Practice*, 7 (4), 411-418.

[26] Walker, L., Philips, J., & Richardson, G. D. (1993, November 10-12). Minority recruitment in teacher education. Paper presented at the Twenty-second annual meeting of the Mid-South Educational Research Association. New Orleans, LA.

Appendix A

Online course evaluation instrument

Please choose the number which best describe your opinion on a scale of 0 to 5, where 0 indicates you strongly disagree with the statement and 5 means you strongly agree with the statement. The first part of the evaluation focuses on the course while the second part focuses on the instructor.

1. I was able to navigate the course web pages with ease.

2. I was able easily to access the course information at the beginning of the course.

3. Course expectations were acceptable and clearly communicated.

4. I liked the way the way course pages were organized.

5. I had to use several resources in this class (e.g., textbook, course presentations, discussions, links, etc.).

6. The web links were relevant.

7. I was able to interact effectively with classmates.

8. I was able to interact effectively with the instructor.

9. I found the discussions useful.

10. I found the course presentations interesting and informative.

11. The use of cooperative learning (if applicable) was well structured.

12. My opinion and input were encouraged and valued.

13. Sharing our research presentations with others in the class was informative.

14. The course assignments were relevant and useful.

15. The course readings were interesting and relevant.

16. The course was intellectually challenging.

17. The course met my educational needs.

18. The instructor was accessible to me by e-mail, phone, or in person.

19. The instructor was well prepared.

20. The instructor posted course assignments on time.

21. The instructor posted grades in a timely fashion.

22. The instructor provided effective feedback on assignments.

23. The instructor maintained a positive atmosphere for learning in the class.

24. The instructor utilized effective teaching methods.

25. The instructor encouraged my participation.

26. The instructor provided relevant topics for discussions.

27. The instructor demonstrated mastery of knowledge of the course materials.

28. The instructor exhibited interest in my learning.

29. Online medium accommodates my learning style.

30. I prefer to have face-to-face classes.