Title: **Improving Course Evaluations to Improve Teaching in Higher Education**

Authors: Theodore Frick, Rajat Chadha, Carol Watson, Pamela Green, Emilija Zlatkovska

Introduction

Course evaluations traditionally used in higher education have few items which are empirically related to student learning achievement. In meta-analyses of studies that have examined this relationship, global items such as "This was an outstanding course." or "The instructor of this course was outstanding." correlate moderately with student achievement (0.40 – 0.50) (e.g., Cohen, 1981; Feldman, 1989; Kulik, 2001). These are currently the *best* predictors of student achievement, and yet do not provide practical feedback on how to improve teaching.

Can course evaluations be improved by incorporating scales that are both valid and useful: student academic learning time (Berliner, 1991), student satisfaction and achievement (Kirkpatrick, 1994), and principles of instruction (Merrill, 2002)? Can course evaluations provide useful feedback on how to improve teaching—in particular, improvements that are associated with increasing the chances of greater student learning achievement?

Academic learning time (ALT) refers to frequency of successful student engagement in learning activities relevant to curriculum goals. ALT has been shown to be positively correlated with student learning achievement (Berliner, 1991; Fisher, Filby, Marliave, Cohen, Dishaw, Moore, & Berliner, 1978).

Kirkpatrick's (1994) four levels of evaluation for gauging training effectiveness have been used for over five decades in non-formal educational settings, such as business and industry. The end-of-term course evaluations typically completed by students are an example of a level 1 (satisfaction) evaluation. Level 2 (learning) evaluation occurs as a result of course assessment tests, course examinations, and other assignments in order to assign a grade for each student.

After an extensive review of the literature, Merrill (2002) synthesized instructional design factors that promote student learning achievement and identified five "first principles" of instruction. Merrill claimed that depending on the extent to which these principles are present during instruction, learning is promoted.

While first principles were drawn from apparently successful instructional theories, few empirical studies have been conducted to verify Merrill's (2002) claim that first principles promote student learning.

Problem

Frick, Chadha, Watson, Wang and Green (2008) piloted a new course evaluation tool (TALQ – Teaching and Learning Quality) consisting of nine scales, including ALT, first principles of instruction, and student ratings of their achievement. Global items were also included in this evaluation.

After analyzing data from 140 respondents spread across 89 different courses, they reported scale reliabilities (Cronbach's α) from 0.74 to 0.97 and statistically significant ($p<0.0005$) moderate to high correlations (0.301 to 0.874) among the scales.

The current study addresses two limitations of the Frick et al. (2008) study:

1. The few respondents from each course might not have been representative of that course. Therefore, the present study sought participation from whole classes and examined the consistency of TALQ scale scores within each participating class.
2. In the first study, learning achievement was self-reported. Therefore, in the present study, an independent measure of student learning achievement, apart from student self reports, was collected from their instructors.

Method

In collaboration with staff from a teaching center, a recruitment email was sent to university faculty that sought volunteers to who were willing to have the TALQ instrument used in their classes, in addition to their normal course evaluations. During last three weeks of the fall 2007 semester, a paper version of the TALQ evaluation was administered by the researchers a week or two before the standard course evaluation used for that class. In two of the classes, TALQ was the only course evaluation instrument that was used, according to that instructor's preferences, since it was administered in the last week of classes.

Items from the TALQ scales were disaggregated and randomly mixed on the course evaluation form, so that students did not know which items belonged to what scale. The course evaluation instrument was administered by the researchers at the beginning of a regular class. Each form had a unique code number on the cover sheet that was repeated on the evaluation form itself. Participating students wrote their names on the top halves of the cover sheets, which were detached and given to the instructor, who then left the classroom. Students completed the TALQ course evaluation anonymously; their individual ratings were collected by the researchers and never shown to instructors.

About one month after completion of the course, instructors rated each participating student's mastery of course objectives using a 10-point scale. Ratings were based on instructor records of grades on student performance in class, completed assignments and projects, exam scores, etc. The bottom halves of the cover sheets with instructor ratings and unique code numbers were returned to the researchers. Thus, student anonymity was maintained, while researchers could pair instructor ratings of student mastery with student ratings of the course by matching the unique code numbers.

Data were collected from 490 students in 12 classes, including business, philosophy, history, kinesiology, social work, computer science and nursing. None of the courses was taught by the researchers. Instructors were provided with summary reports of TALQ scales after they had submitted their ratings of student mastery of course objectives, except for one instructor who needed the reports for her annual report and these were the sole course evaluations she had used.

Results

Each of the nine TALQ scales had 3 to 5 items, and data analysis indicated that reliabilities of the scales were generally very high. Cronbach alpha coefficients ranged from 0.77 to 0.94 except the authentic problems scale which was 0.59. When the five scales for first principles of instruction were combined, the alpha coefficient was 0.87.

We found statistically significant Spearman rho correlations ($p < 0.0005$) between first principles of instruction and global ratings of course and instructor quality ($r = 0.748$), reported student satisfaction ($r = 0.774$,), reported ALT ($r = 0.574$), and student perceptions of their learning progress ($r = 0.720$). Additionally, ALT was statistically significantly correlated with global ratings ($r = 0.525$), satisfaction ($r = 0.558$), learning progress ($r = 0.505$), and instructor rating of student mastery ($r = 0.394$). Instructor ratings of student mastery and student perceptions of their mastery of course objectives were also significantly correlated ($r = 0.43$).

Analysis of patterns in time (Frick, 1990) was used to further investigate these relationships. When students agreed that first principles of instruction occurred, they were about 3 times more likely to agree that ALT also occurred, compared to when they did not agree that first principles occurred. APT results indicated that students were nearly 4 times more likely to achieve high mastery of course objectives (according to their instructor's ratings), when those students agreed that *both* first principles of instruction and ALT occurred in their courses, compared to course evaluation ratings from students who did not agree that both occurred.

Conclusions

First principles of instruction are something that college instructors can control. Academic learning time (ALT) is under the control of the student. ALT is indicated by student effort to learn, effort that results in frequently performing successfully in course activities. Data from this study indicate that when ALT and first principles were both reported to occur, the likelihood of a high level of student mastery of course objectives (according to instructor evaluation of student performance) is about 4 times greater than the likelihood of high mastery when neither first principles nor ALT were reported to occur.

Low course evaluation scores on global items do not tell instructors how to improve their teaching in ways that are likely to improve student learning achievement.

On the other hand, the TALQ scales on first principles of instruction can be used to identify areas in which teaching and course design can be improved—i.e., instructors can incorporate authentic, real-world problems for students to solve, activate student learning, demonstrate what is to be learned, provide opportunities for students to successfully solve problems with coaching and feedback, and help students integrate what they have learned into their personal lives. Incorporation of these first principles of instruction in courses is strongly associated with high ratings of academic learning time, student satisfaction, instructor ratings of student learning achievement and student ratings of overall instructor and course quality.

References

Berliner, D. (1991). What's all the fuss about instructional time? In M. Ben-Peretz & R. Bromme (Eds.), *The nature of time in schools: Theoretical concepts, practitioner perceptions*. New York: Teachers College Press.

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

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30*, 583–645.

Fisher, C., Filby, N., Marliave, R., Cohen, L., Dishaw, M., Moore, J., & Berliner, D. (1978). *Teaching behaviors: Academic Learning Time and student achievement: Final report of Phase III-B, Beginning Teacher Evaluation Study*. San Francisco: Far West Laboratory for Educational Research and Development.

Frick, T. (1990). Analysis of patterns in time (APT): A method of recording and quantifying temporal relations in education. *American Educational Research Journal, 27*(1), 180-204.

Frick, T.W. , Chadha, R. , Watson, C. , Wang, Y. , Green, P. (2008, in press). College student perceptions of teaching and learning quality. *Educational Technology Research and Development*.

Kirkpatrick, D. (1994). *Evaluating Training Programs: The Four Levels.* San Francisco, CA: Berrett-Koehler.

Kulik, J. A. (2001). Student ratings: Validity, utility and controversy. *New Directions for Institutional Research, 109*, 9-25.

Merrill, M. D. (2002). First principles of instruction. *Education Technology Research & Development, 50*(3), 43-59.