Theory-Based Evaluation of Instruction: Implications for Improving Student Learning Achievement in Postsecondary Education

Theodore Frick, Rajat Chadha, Carol Watson, and Ying Wang

Abstract While student global ratings of college courses historically predict learning achievement, the majority of recent U.S. college graduates lack proficiency in desired skills. Teaching and Learning Quality (TALQ), a new course evaluation instrument, was developed from extant instructional theory that promotes student learning. A survey of 193 students in 111 different courses at multiple institutions was conducted using TALQ. Results indicated strong associations among student ratings of First Principles of Instruction, academic learning time, perceptions of learning gains, satisfaction with courses, perceived mastery of course objectives, and their overall evaluation of courses and instructors. Instructors can implement the theoretically derived First Principles of Instruction by challenging students with real-world problems or tasks, activating student learning, demonstrating what is to be learned, providing feedback on student learning attempts, and encouraging student integration of learning into their personal lives.

T. Frick (✉)
Department of Instructional Systems Technology, School of Education, Indiana University, Bloomington, IN, USA
e-mail: frick@indiana.edu

R. Chadha (✉)
Department of Instructional Systems Technology, School of Education, Indiana University, Bloomington, IN, USA
e-mail: rajatchadha@gmail.com

C. Watson (✉)
Eppley Institute for Parks and Public Lands, Indiana University, Bloomington, IN, USA
e-mail: watsonc@indiana.edu

Y. Wang (✉)
Department of Education, Northwestern College, St. Paul, MN, USA
e-mail: ywang@nwc.edu

M. Orey et al. (eds.), *Educational Media and Technology Yearbook*,
Educational Media and Technology Yearbook 35,
DOI 10.1007/978-1-4419-1516-0_5,

Keywords Teaching and learning quality · Higher education · Student learning · Course evaluation · First Principles of Instruction · Academic learning time

Problem

This study began because the first author served on a university committee that was expected to choose a few outstanding college instructors as recipients of significant monetary awards. The top candidates recommended by their departments had provided the committee with customary forms of evidence that have been used for evaluation of teaching for promotion and tenure. This experience nonetheless raised the question: What empirical evidence is there that course evaluation data are associated with student learning achievement?

Thus, we began to review research on student course evaluation in higher education. A review by Cohen (1981) stood out as the most highly cited in the *Web of Knowledge* by scholarly research studies subsequently published on this issue. Cohen's study

> . . . used meta-analytic methodology to synthesize research on the relationship between student ratings of instruction and student achievement. The data for the meta-analysis came from 41 independent validity studies reporting on 68 separate multisection courses relating student ratings to student achievement. The average correlation between an overall instructor rating and student achievement was 0.43; the average overall course rating and student achievement was 0.47. . . . The results of the meta-analysis provide strong support for the validity of student ratings as measures of teaching effectiveness. (p. 281)

According to Cohen (1981), a typical example of an overall instructor rating item was "The instructor is an excellent teacher." A typical overall course rating item was "This is an excellent course." Cohen also found that ratings of instructor *skill* correlated on average 0.50 with student achievement (e.g., "The instructor has good command of the subject matter," "The instructor gives clear explanations"). The other factor that showed a high average correlation (0.47) was course *structure* (e.g., "The instructor has everything going according to course schedule," "The instructor uses class time well").

Studies similar to Cohen's meta-analysis have since been conducted, and those that are methodologically sound have yielded relatively consistent findings (Abrami, d'Apollonia, & Cohen, 1990; Abrami, 2001; Feldman, 1989; Kulik, 2001; Marsh, 1984). Further studies have also demonstrated positive relationships between independently observed classroom behaviors and student ratings of instructors and courses (cf. Koon & Murray, 1995; Renaud & Murray, 2004). When these studies are taken as a whole, reported correlations are moderate and positive, typically in the 0.30–0.50 range. At first glance, there appears to be little doubt that at least global student ratings of instructors and courses predict student achievement in higher education.

However, such ratings are at best moderately or weakly correlated with student learning achievement – explaining a relatively small proportion of variance

in student learning achievement (Emery, Kramer, & Tian, 2003). In a more recent example, Arthur, Tubré, Paul, and Edens (2003) conducted a pre-/post-study of student learning gains in an introductory psychology course. They found a *weak* relationship between student evaluations of teaching effectiveness and measures of student learning gains. They also reported a *moderate* relationship between student grades and learning achievement.

Another potentially confounding factor is that students may respond to course evaluations in ways that do not reflect course or instructor quality. For example, Clayson, Frost, and Sheffet (2006) empirically tested the "reciprocity effect" between student grades and their ratings of instructors and classes. They found that when grades were lowered within a class, the ratings decreased, and when grades were raised, ratings increased. Clayson et al. (2006) offered the hypothesis that ". . .students reward instructors who give them good grades and punish instructors who give them poor grades, irrespective of any instructor or preexisting student characteristic" (p. 52).

Recent Reports on College Student Achievement – or Lack Thereof

Perhaps the issue of course evaluation should be further examined in light of what appears to be unsatisfactory levels of student achievement in postsecondary education. Two recent reports were studied in more detail. In the first report, Baer, Cook, and Baldi (2006) assessed literacy skills of 1,827 students who were nearing completion of their degrees at 80 randomly selected 2- and 4-year public universities and colleges. They used the same standardized assessment instrument as that in the National Assessment of Adult Literacy. The literacy assessments were supervised by a test administrator on each campus.

The Baer et al. (2006) report provides some sobering findings. They reported percentages of students from 2-year versus 4-year institutions, respectively, 23 and 38% of whom were *proficient* in prose literacy, 23 and 40% in document literacy, and 18 and 34% in quantitative literacy. This means that more than 75% of students at 2-year institutions performed *lower than proficiency level*, and more than 50% at 4-year institutions likewise scored lower. For example, these students could *not* "perform complex literacy tasks, such as comparing credit card offers with different interest rates or summarizing the arguments of newspaper editorials" (American Institutes for Research, 2006, n.p.). Even worse,

> . . .approximately 30 percent of students in 2-year institutions and nearly 20 percent of students in 4-year institutions have only Basic quantitative literacy. Basic skills are those necessary to compare ticket prices or calculate the cost of a sandwich and a salad from a menu. (American Institutes for Research, 2006, n.p.)

In the second report, a comprehensive review of the literature by Kuh, Kinzie, Buckley, Bridges, and Hayek (2006) indicated a number of factors that influence student success in postsecondary education. One of their major findings was: "[a]mong the institutional conditions linked to persistence are supportive peers, faculty and staff members who set high expectations for student performance, and academic

programs and experiences that actively engage students and foster academic and social integration" (p. 4). Based on these and other findings, Kuh et al., (2006) made several recommendations. One important recommendation was to "*. . . focus assessment and accountability efforts on what matters to student success*" (p. 4, italics added).

Research Questions

Results from these recent studies provide impetus for reexamining the kinds of items used on typical course evaluations in higher education. This led us to ask the primary research questions addressed in this report: Can we develop reliable scales for course evaluation that measure factors that are supported by instructional theory? Do these scales identify how instruction might be improved in ways that are more likely to be associated with improved student learning and overall course quality?

If we *can* develop better scales for use in course evaluation, then this would address, in part, the important recommendation made by Kuh et al. (2006) that universities and colleges should focus their assessment efforts on factors that influence student success. Course evaluations could be one of those assessments.

First Principles of Instruction. After an extensive review of the literature on theories and models of instruction, Merrill (2002) synthesized factors that promote student learning achievement. He identified what he called "First Principles" of Instruction. He claimed that to the extent these principles are present during instruction, learning is promoted. These First Principles include (1) *authentic problems or tasks* (students engage in a series of increasingly complex real-world problems or authentic whole tasks); (2) *activation* (students engage in activities that help them link past learning or experience with what is to be newly learned); (3) *demonstration* (students are exposed to differentiated examples of what they are expected to learn or do); (4) *application* (students solve problems or perform whole tasks themselves with scaffolding and feedback from instructors or peers); and (5) *integration* (students engage in activities that encourage them to incorporate what they have learned into their own personal lives). Instructors can do something about First Principles of Instruction in their courses. If instructors use more of the First Principles in their teaching, instructional theory predicts that students should learn more. First Principles of Instruction are not specific to a particular subject matter content, according to Merrill, and thus have a wide range of applicability.

Academic learning time. In examining the research literature, one factor has consistently shown a strong relation to student achievement at all levels: academic learning time (ALT). ALT refers to the frequency and amount of time that students spend *successfully engaged in learning tasks* that are similar to skills and knowledge they will be later tested on (Berliner, 1991; Brown & Saks, 1986; Fisher et al., 1978; Kuh et al., 2006; Squires, Huitt, & Segars, 1983). Yet the kinds of items in the Cohen (1981) meta-analysis largely focused on the instructor or course, not on *student* ALT. Student ALT is not something an instructor has direct control over, since it is the students who must put in the effort to succeed on tasks and activities

in the course. However, if instruction is more effective, then one indicator of this effectiveness would be higher levels of ALT. Thus, a high rating of ALT by the students of their own performance would be an indicator of a successful course. Since ALT is predictive of student learning achievement, increased use of First Principles of Instruction would be expected to result in increased student ALT.

Levels of evaluation of training. Finally, we considered levels of evaluation of training effectiveness that have been used for more than five decades in nonformal educational settings such as business and industry (Kirkpatrick, 1994). The four levels of evaluation are (1) learner *satisfaction* with the training, often referred to as a "smiles test" or reaction; (2) *learning achievement*; (3) *transfer* of learning to the learner's job or workplace[1]; and (4) *impact* on the overall organization to which the learner belongs.

Level 1 is what many people believe that traditional course evaluations often measure, i.e., student satisfaction with a course and instructor. If a course and instructor is good, from the perspective of a student, then he or she would be expected to be more satisfied as a result.

With respect to Level 2, student learning achievement, we wondered if we could get students to rate their own learning progress. That is, compared with what they knew or could do before they took the course, how much did they perceive that they had learned? While there are issues of validity of self-reports, Cohen (1981) and Kulik (2001) indicated that many studies have found positive correlations of such self-reports with objective assessments in college such as common exams in multisection courses. Learning progress would be a desirable outcome of a course, just as student ALT and satisfaction. Learning progress is nonetheless not a measure of actual student learning achievement, but only a perception by students about how much they have learned.

One might expect course grades to indicate student learning achievement in a more objective manner. However, with apparent grade inflation these days, course grades are probably not a good indicator of student learning achievement. Nonetheless, we wondered how students perceived their mastery of course objectives. It would be possible for students to report that they had learned a great deal in a course, but nonetheless they had not mastered the course objectives. Indeed, how to measure Kirkpatrick's Level 2 is somewhat elusive, particularly if instructor grades of student performance are not valid indicators of student learning achievement. Independent measures of student skills and knowledge are needed. Attempts to measure college student prose literacy, document literacy, and quantitative literacy, such as the study by Baer et al. (2006), would be an example of an independent measure of student achievement in college. However, we do not have standardized assessments of student learning at the university level in general in the United States, although some professions have their own tests as part of licensing or certification, such as for medical practitioners, optometrists, and lawyers.

[1]It should be also noted that Kirkpatrick's Level 3 is highly similar to Merrill's Principle 5 (integration). We did not attempt to measure Level 4 in this study.

Method

A survey instrument was constructed that contained items intended to measure scales for student ratings of self-reported ALT, satisfaction with the course, learning progress, authentic problems, activation, demonstration, application, and integration. In addition, several items were included from the university's standard course evaluation item pool from the Bureau for Evaluative Studies and Testing (BEST). These BEST items included *global* ones similar to those reported in Cohen (1981), which indicated overall ratings of the course and instructor. We also included on the survey several demographic questions, the grade that they received or expected to receive in the course, and their rating of their mastery of course objectives.

See Table 2 for the nine a priori item sets. Each set contained five items intended to measure the respective construct (scale). For this study, five items per scale were used with the anticipation that reliability analysis would permit scale reduction without compromising internal consistency reliability.

A paper version of the instrument was then reviewed by several faculty instructors, and wording of items considered to be confusing or ambiguous was modified. The instrument, now referred to as the *Teaching and Learning Quality Scales* (*TALQ Scales*), was then converted to a Web survey, which can be viewed online at http://www.indiana.edu/~edsurvey/evaluate/.

No explicit reference was made to Merrill's First Principles of Instruction or Kirkpatrick's levels of evaluation in the survey or study information sheet. Student ratings were not shared with their instructors and hence could not affect their grade in the course.

Volunteers were sought for participation in the study through e-mail requests to faculty distribution lists and student organizations at several postsecondary institutions. Respondents who had nearly or recently completed a course completed the survey. There were 193 valid cases remaining after elimination of those containing no data or that were test cases to ensure that data collection was working as intended via the Web survey.

Results

Since participation in the survey was voluntary, we also collected demographic data in the survey in order to facilitate interpretation of results and to document the representativeness of the obtained sample of 193 cases.

Nature of Courses and Respondents

Course topics. Data indicated that respondents evaluated a wide range of courses with relatively few respondents from any given course. We conducted a content analysis of qualitative responses to the survey question about the course title or content. A total of 111 different subject areas were mentioned by 174 respondents (19 respondents did not answer this question).

While courses in business (34), medicine (23), education (18), English (18), and computers and technology (12) were mentioned more frequently than others, a very wide range of subject matter was represented in the courses taken by respondents. Thus, there were 111 courses that appeared to have unique subject matter or titles, and the remaining 63 had similar course titles as mentioned by at least one other respondent (though seldom with the same instructor).

Course instructors. In addition, content analysis of courses rated by students indicated that they were, by and large, taught by different instructors. While several instructor names with the same or approximate spellings were listed more than once by different respondents, the very large majority of respondents appeared to have different instructors. This is consistent with the wide range of course topics, as indicated above.

Gender of student respondents. In Table 1, it can be seen that 132 females and 55 males responded to the survey (6 did not report gender). While it may appear that a disproportionate number of females responded, for the scales investigated in this

Table 1 Respondent and course demographics ($N = 193$)

Question		Frequency	Percentage
Gender	Female	132	70.6
	Male	55	29.4
	Missing	6	3.1
Class rating: I would rate this class as:	Great	107	56.0
	Average	71	37.2
	Awful	13	6.8
	Missing	2	1.0
Expected grade: In this course, I expect to receive (or did receive) a grade of:	A	116	64.1
	B	52	28.7
	C	11	6.1
	D	2	1.1
	Missing	12	6.2
Achievement: With respect to achievement of objectives of this course, I consider myself a:	Master	44	22.9
	Partial master	117	60.9
	Nonmaster	31	16.1
	Missing	1	0.5
Class standing: I am a:	Freshman	32	17.4
	Sophomore	25	13.6
	Junior	38	20.7
	Senior	30	16.3
	Graduate	59	32.1
	Missing/other	9	4.7
Course setting: I took this course:	Face-to-face	116	60.4
	Blended	12	6.3
	Online	64	33.3
	Missing	1	0.5

study, there were *no* significant relationships between gender and other variables or scales, as discussed below.

Class standing of respondents. In Table 1, it can be seen that approximately one-third of respondents were graduate students and the remaining two-thirds were undergraduates, with the latter being distributed about equally among freshmen to seniors (14–21% in each group).

Course settings. About 60% of courses evaluated were face-to-face, and about one-third were online or distance courses.

Course grades. Table 1 also displays responses of students with respect to their course grade. Almost 93% reported that they received or expected to receive an A or a B.

Mastery of course objectives by students. Since grades were not anticipated by this research team to be very discriminating among respondents, they were also asked, "With respect to achievement of objectives of this course, I consider myself a ____." Choices were master, partial master, and nonmaster. Table 1 indicates that about 23% reported themselves to be masters. The large majority considered themselves to be partial masters of course objectives, while 16% identified themselves as nonmasters.

Relationships Among Variables

In this study, we choose our a priori Type I error rate as $\alpha = 0.0005$ for determining statistical significance. Our sample size was fairly large ($n = 193$ cases), and we sought to minimize the probability of concluding statistical significance as an artifact of numerous comparisons. We conducted a total of 58 statistical tests. The overall Type I error rate for this study was $1 - (1 - 0.0005)^{58} = 0.0286$ (cf. Kirk, 1995, p. 120).

Gender. Gender (1 = male, 0 = female) was not significantly related ($p > 0.0005$) to overall course rating,[2] expected or received grade,[3] mastery level,[4] or class standing.[5] One of the chi squares approached significance ($\chi^2 = 5.22$, df = 2, $p = 0.052$, $n = 189$) between gender and mastery level. Slightly more males considered themselves to be masters than expected, and slightly fewer females considered themselves as masters than expected if there were no relationship.

One-way ANOVAs were run between gender and each of the remaining scales and variables discussed below. None of the *F*s was statistically significant.

Student mastery level. Spearman's ρ indicated a significant association between class rating and mastery of course objectives ($\rho = 0.306$, $p < 0.0005$, $n = 191$).

---

[2] 2 = great, 1 = average, 0 = awful

[3] 4 = A, 3 = B, 2 = C, 1 = D, 0 = F

[4] 2 = master, 1 = partial master, 0 = nonmaster,

[5] 5 = graduate, 4 = senior, 3 = junior, 2 = sophomore, 1 = freshman

Students who considered themselves masters of course objectives were more likely to rate the course as "great." There was also a significant correlation between student reports of mastery level and course grades ($\rho = 0.397$, $p < 0.0005$, $n = 181$).

Grades. Students' expected or received course grades were weakly associated with their ranks of overall course quality ($\rho = 0.241$, $p = 0.001$, $n = 180$). Grades and class standing were also weakly related ($\rho = 0.230$, $p = 0.002$, $n = 174$). Graduate students and upperclassmen reported somewhat higher grades than freshmen and sophomores.

Scale Reliabilities

Scale items and reliabilities are listed in Table 2. To determine the reliability of each scale, all five items in each scale were initially used to compute internal consistency with Cronbach's α coefficient. Items that were negatively worded (–) had their Likert scores reversed. Items were removed until no further item could be removed without decreasing the α coefficient. It should be noted that factor analysis was not considered appropriate at this point, since these scales were formed a priori.

Our goal was to form a single scale score for each reliable scale before further analysis of relationships among variables measured in the study. It can be seen in Tables 2 and 3 that internal consistency of each scale was generally quite high, ranging from 0.81 to 0.97.

Combined First Principles scale (Merrill 1–5). To determine the reliability of the combined scale, we first formed a scale score for each First Principle by computing a mean rating score for each case. Then we entered the five First Principles scale

Table 2 Nine TALQ Scales

Item no.	Scale name, Cronbach alpha, and items stems for each scale[6]
1.	*Academic Learning Time Scale ($\alpha = 0.81$)*
1-	I did not do very well on most of the tasks in this course, according to my instructor's judgment of the quality of my work
12	I frequently did very good work on projects, assignments, problems and/or learning activities for this course
14	I spent a lot of time doing tasks, projects and/or assignments, and my instructor judged my work as high quality
24	I put a great deal of effort and time into this course, and it has paid off – I believe that I have done very well overall
29-	I did a minimum amount of work and made little effort in this course

[6] Item numbers followed by a minus are negatively worded, and scales were reversed for reliability analyses.

Table 2 (continued)

Item no.	Scale name, Cronbach alpha, and items stems for each scale
2.	*Learning Progress Scale ($\alpha = 0.95$)*
4	Compared to what I knew before I took this course, I learned a lot
10	I learned a lot in this course
22	Looking back to when this course began, I have made a big improvement in my skills and knowledge in this subject
27-	I learned very little in this course
32-	I did not learn much as a result of taking this course
3.	*Global rating items selected from the standard university form ($\alpha = 0.97$)*
8	Overall, I would rate the quality of this course as outstanding
16	Overall, I would rate this instructor as outstanding
38	Overall, I would recommend this instructor to others
4.	*Authentic Problems/Tasks Scale ($\alpha = 0.87$)*
3	I performed a series of increasingly complex authentic tasks in this course
19	My instructor directly compared problems or tasks that we did, so that I could see how they were similar or different
25	I solved authentic problems or completed authentic tasks in this course
31	In this course I solved a variety of authentic problems that were organized from simple to complex
33	Assignments, tasks, or problems I did in this course are clearly relevant to my professional goals or field of work
5.	*Activation Scale ($\alpha = 0.90$)*
9	I engaged in experiences that subsequently helped me learn ideas or skills that were new and unfamiliar to me
21	In this course I was able to recall, describe or apply my past experience so that I could connect it to what I was expected to learn
30	My instructor provided a learning structure that helped me to mentally organize new knowledge and skills
39	In this course I was able to connect my past experience to new ideas and skills I was learning
41-	In this course I was not able to draw upon my past experience nor relate it to new things I was learning
6.	*Demonstration Scale ($\alpha = 0.89$)*
5	My instructor demonstrated skills I was expected to learn in this course
17	My instructor gave examples and counter-examples of concepts that I was expected to learn
35-	My instructor did not demonstrate skills I was expected to learn
43	My instructor provided alternative ways of understanding the same ideas or skills
7.	*Application Scale ($\alpha = 0.82$)*
7	My instructor detected and corrected errors I was making when solving problems, doing learning tasks or completing assignments

Table 2 (continued)

Item no.	Scale name, Cronbach alpha, and items stems for each scale
36	I had opportunities to practice or try out what I learned in this course
42	My course instructor gave me personal feedback or appropriate coaching on what I was trying to learn
8.	*Integration Scale ($\alpha = 0.87$)*
11	I had opportunities in this course to explore how I could personally use what I have learned
28	I see how I can apply what I learned in this course to real life situations
34	I was able to publicly demonstrate to others what I learned in this course
37	In this course I was able to reflect on, discuss with others, and defend what I learned
44-	I do not expect to apply what I learned in this course to my chosen profession or field of work
9.	*Learner Satisfaction Scale ($\alpha = 0.94$)*
2	I am very satisfied with how my instructor taught this class
6-	I am dissatisfied with this course
20-	This course was a waste of time and money
45	I am very satisfied with this course

Table 3 Combined First Principles Scale ($\alpha = 0.94$)

Principle
Authentic Problems/Tasks: students engage in real-world problems and tasks or activities
Activation: student prior learning or experience is connected to what is to be newly learned
Demonstration: students are exposed to examples of what they are expected to learn or do
Application: students try out what they have learned with instructor coaching or feedback
Integration: students incorporate what they have learned into their own personal lives

scores into the reliability analysis, treating each principle score as an item score itself. The resulting Cronbach's α coefficient was 0.94.

Formation of remaining scale scores. Scores were created for remaining scales such that each scale score represented a mean Likert score for each case.

Correlational Analyses

We next investigated the relationships among the scales themselves. Spearman's ρ was used as a measure of association, since these scales are ordinal. The reader should note that we considered a correlation to be significant when $p < 0.0005$, based on Type I error rate for this study, which in effect means that a finding was considered statistically significant when $p < 0.0286$.

First Principles of Instruction considered individually. It can be seen in Table 4 that First Principles are highly correlated with each other, with all correlations

Table 4 Spearman's ρ correlations for First Principles of Instruction scales

		Authentic problems	Activation	Demonstration	Application	Integration
Authentic Problems Scale	*P*	1.000				
	N	192				
Activation Scale	ρ	0.790[a]	1.000			
	N	192	193			
Demonstration Scale	ρ	0.803[a]	0.792[a]	1.000		
	N	189	190	190		
Application Scale	ρ	0.724[a]	0.763[a]	0.794[a]	1.000	
	N	186	186	184	186	
Integration Scale	ρ	0.819[a]	0.818[a]	0.770[a]	0.722[a]	1.000
	N	192	193	190	186	193

[a] Correlation is significant ($p < 0.0005$, 2-tailed).

significant at $p < 0.0005$, with ρ ranging from 0.722 to 0.819. This should not be surprising, since the internal consistency α is 0.94. Therefore, the five First Principles were combined into a single scale score, as described above for subsequent analyses.

Relationships among scales. The results in Table 5 are very strong as a group. Except for student mastery, the Spearman correlations ranged from 0.46 to 0.89, with most in the range 0.60–0.80. Students who agreed that they frequently engaged successfully in problems and doing learning tasks in a course (reported ALT) also were more likely to report that they mastered course objectives. Furthermore, they

Table 5 Spearman's ρ correlations among TALQ Scales

		Combined First Principles	ALT	Learning progress	Satisfaction	Global rating (BEST)	Class rating	Mastery
Combined First Principles	ρ	1.000						
	N	193						
ALT	ρ	0.670[a]	1.000					
	N	192	192					
Learning Progress	ρ	0.833[a]	0.747[a]	1.000				
	N	193	192	193				
Satisfaction	ρ	0.850[a]	0.683[a]	0.856[a]	1.000			
	N	192	191	192	192			
Global Rating (BEST)	ρ	0.890[a]	0.605[a]	0.811[a]	0.903[a]	1.000		
	N	193	192	193	192	193		
Class Rating	ρ	0.694[a]	0.464[a]	0.649[a]	0.753[a]	0.773[a]	1.000	
	N	191	190	191	190	191	191	
Mastery of Objectives	ρ	0.344[a]	0.359[a]	0.334[a]	0.317[a]	0.341[a]	0.306[a]	1.000
	N	192	191	192	191	192	191	192

[a] Correlation is significant ($p < 0.0005$, 2-tailed).

agreed that this was an excellent course and instructor, and they were very satisfied with it.

There were strong relationships between ALT and First Principles of Instruction. Students who agreed that First Principles were used in the course also agreed that they were frequently engaged successfully in solving problems and doing learning tasks. These relationships will be clarified in the pattern analysis results described below (analysis of patterns in time [APT]).

Pattern Analysis (APT)

While there were numerous highly significant bivariate relationships that explained typically between 40% and 80% of the variance in ranks, specific patterns that show temporal relations among three or more variables are not shown in Tables 4 and 5. For example, what is the likelihood that *if* students agreed that ALT occurred during the course, *and if* they also agreed that First Principles occurred during the course, *then* what is the likelihood that they agreed that they learned a lot in the course?

Analysis of patterns in time (APT) is one way of approaching data analysis that is an alternative to the linear models approach (e.g., regression analysis, path analysis, and ANOVA; see Frick, 1983, 1990; Frick, An, & Koh, 2006):

> This [APT] is a paradigm shift in thinking for quantitative methodologists steeped in the linear models tradition and the measurement theory it depends on (cf. Kuhn, 1962). The fundamental difference is that *the linear models approach relates independent measures through a mathematical function and treats deviation as error variance*. *On the other hand, APT measures a relation directly by counting occurrences of when a temporal pattern is true or false in observational data*. Linear models relate the measures; APT measures the relation. (Frick et al., 2006, p. 2)

In the present study, we wanted to know that if students reported that ALT and First Principles occurred, then what is the likelihood that students also reported that they learned a lot, mastered course objectives, or were satisfied with their instruction?

We were able to do APT with our data set as follows: New dichotomous variables from existing scale scores were created for each of the cases.[7] A scale was recoded as "Yes" if the scale score for that case was greater than or equal to 3.5, and "No" if less than 3.5. For example, if the ALT agreement code is "Yes," it means that the student "agreed" or "strongly agreed" that ALT occurred for him or her in that course (frequent, successful engagement in problems, tasks, or assignments); and if the code is "No," then the student did *not* "agree" or "strongly agree" that ALT occurred for him or her.

---

[7] Variables can be characterized by more than two categories, but for this study and the sample size and the numbers of combinations, a simple dichotomy appeared to be best – especially since ratings were negatively skewed.

Table 6 APT Frequencies for the pattern: If *ALT* and *First Principles,* then *learning progress?*

	ALT agreement			
	No		Yes	
	Combined First Principles Agreement		Combined First Principles Agreement	
	No	Yes	No	Yes
	Learning progress agreement	Learning progress agreement	Learning progress agreement	Learning progress agreement
	Count	Count	Count	Count
No	26	8	10	6
Yes	9	8	12	113

If *ALT* and *First Principles*, then *Learned a Lot*. InTable 6 results are presented for the APT query. If student agreement with ALT is Yes, and if student agreement with First Principles is Yes, then student agreement with Learned a Lot is Yes? Normally in APT one would have a number of observations *within* a case for a temporal pattern, so that a probability can be calculated for each case and the probabilities averaged across cases. For example, in the Frick (1990) study, probabilities of temporal patterns on each case were determined from about 500 time samples. In the present study, we have only one observation per classification (variable) for each case.

There were a total of 119 occurrences of the antecedent condition (if student agreement with ALT is Yes, *and* if student agreement with First Principles is Yes). Given that the antecedent was true, the consequent (student agreement with Learned a Lot is Yes) was true in 113 out of those 119 cases, which yields an APT conditional probability estimate of 113/119 or 0.95 for this pattern.

Next we investigated the pattern: If student agreement with ALT is No, and if student agreement with First Principles is No, then student agreement with Learned a Lot is Yes? It can be seen that the antecedent occurred a total of 35 times, and the consequent occurred in 9 out of those 35 cases, for a conditional probability estimate of $9/35 = 0.26$. Thus, about 1 out of 4 students agreed that they learned a lot in the course when they did not agree that ALT and First Principles occurred.

This can be further interpreted: When both ALT and First Principles occurred, students were nearly four times as likely ($0.95/0.26 = 3.7$) to agree that they learned a lot in the course, compared to when ALT and First Principles are reported to not occur.

If *ALT* and *First Principles*, then *Learner Satisfaction*. In Table 7, results for the APT query are presented: If student agreement with ALT is Yes, and if student agreement with First Principles is Yes, then student agreement with Learner Satisfaction is Yes? The consequent was true in 113 out of 118 cases when the

Table 7 APT Frequencies for the pattern: If *ALT* and *First Principles,* then *learner satisfaction?*

	ALT agreement			
	No		Yes	
	Combined First Principles agreement		Combined First Principles agreement	
	No	Yes	No	Yes
	Satisfaction agreement	Satisfaction agreement	Satisfaction agreement	Satisfaction agreement
	Count	Count	Count	Count
No	25	6	11	5
Yes	10	10	11	113

antecedent was true for a probability estimate of 0.96. On the other hand, when ALT was No and First Principles was No, then Learner Satisfaction occurred in 10 out of 35 cases, or a probability estimate of 0.29. The estimated odds of Learner Satisfaction when both ALT and First Principles are present compared to when both are not are about 3.3–1 (0.96/0.29).

If *ALT* and *First Principles*, then *Outstanding Instructor/Course*. In Table 8, results for the APT query are presented: If student agreement with ALT is Yes, and if student agreement with First Principles is Yes, then student agreement with Outstanding Instructor/Course is Yes? The probability of this pattern is $114/119 = 0.96$. If both antecedent conditions are false, the probability is $4/35 = 0.11$. The odds are about 8.7–1 that an instructor/course is viewed as outstanding by students

Table 8 APT Frequencies for the pattern: If *ALT* and *First Principles*, then outstanding instructor/course (global rating)?

	ALT agreement			
	No		Yes	
	Combined First Principles agreement		Combined First Principles agreement	
	No	Yes	No	Yes
	Global rating agreement	Global rating agreement	Global rating agreement	Global rating agreement
	Count	Count	Count	Count
No	31	4	15	5
Yes	4	12	7	114

Table 9 APT Frequencies for the pattern: If *ALT* and *First Principles,* then *mastery of course objectives?*

	ALT agreement			
	No		Yes	
	Combined First Principles agreement		Combined First Principles agreement	
	No	Yes	No	Yes
	Mastery level	Mastery level	Mastery level	Mastery level
	Count	Count	Count	Count
Nonmastery	14	3	3	11
Partial mastery	19	9	15	73
Mastery	2	4	4	34

when ALT and First Principles are both present versus both absent, according to student ratings.

If *ALT* and *First Principles*, then *Mastery*. In Table 9 results for the APT query are presented: If student agreement with ALT is Yes, and if student agreement with First Principles is Yes, then student agreement with Mastery is Yes? Here the pattern is less predictable, since it was true for 34 out of 118 students for a probability of 0.29 (roughly 1 out of 3 students). On the other hand, only 2 out of 35 students agreed that they had mastered course objectives (probability = 2/25 = 0.06) when they did not agree that First Principles and ALT occurred. Thus, students were five times more likely to agree that they mastered course objectives when they agreed versus did not agree that both ALT and First Principles occurred when they took the course.

Discussion

Implications from APT findings. The APT findings are consistent with earlier correlational results. APT allows temporal combinations or patterns of more than two variables at a time. In APT, relationships are not assumed to be linear nor modeled by a mathematical function – e.g., as in regression analysis. APT probability estimates are relatively easy to comprehend and can have practical implications. The reader is cautioned that a temporal association does not imply causation (cf. Frick, 1990).

Mastery of learning objectives. As noted earlier, less than 1 out of 4 students considered themselves masters of course objectives, even though 93% received As and Bs for their course grades. This could be interpreted in a number of ways, but what is noteworthy is the large discrepancy between grades received and student perceptions of their mastery. While student grades and perceptions of mastery are

significantly correlated ($\rho = 0.397$), a grade of A or B appears not to be a good indicator of mastery of course objectives. A cross-tabulation of grades by mastery level indicated that 39 out of 182 students (21.4%) considered themselves to be masters and who received grade A. Approximately 42% of all students received an A, who perceived themselves to be partial masters (37%) or nonmasters (5%) of course objectives.

Implications from First Principles of Instruction. We did not tell students that we were measuring First Principles. We constructed rating scale items that were consistent with each of the five First Principles; then we scrambled the order and mixed them with items measuring other scales on the survey. Data from our study indicate that these rating scales are highly reliable.

While further research is needed with respect to the validity of the scales, those scales that rate use of First Principles of Instruction reveal things that course instructors can do something about. For example, if scores on the authentic problems/task scale are low, instructors could consider revising their course so that students are expected to perform authentic problems or tasks as part of their learning. If scores on the integration scale are low, then new activities can be included in a course to encourage students to incorporate what they have learned in their real lives. In other words, such changes would make course objectives more relevant from a student's perspective. If learning activities are viewed as being more relevant, then students would be expected to be more motivated and to spend more time engaged in activities than before. More successful engagement should lead to greater achievement, according to past studies of ALT (e.g., see Kuh et al., 2006). It is very clear from results in this study that students who agree that First Principles were used in their courses are also likely to agree that such courses and instructors were outstanding ($\rho = 0.89$).

The reader should note that numerous studies in the past have shown significant positive correlations between global course ratings and objective measures of student achievement such as course exams in multiple sections (Cohen, 1981; Kulik, 2001). Thus, it is likely that use of First Principles of Instruction is correlated with student learning achievement, but that was not measured in this study. It is important to note, however, in a separate study of undergraduate students in 12 courses at one university (Frick, Chadha, Watson, & Zlatkovska, 2009), the TALQ Scales were compared with independent assessments by classroom instructors of each student's mastery of course objectives. In that study, the TALQ was completed by most students in each of those 12 courses (total $n = 464$), and similar patterns of results were found. For example, students who agreed that their instructors used First Principles of Instruction were nearly three times more likely to agree that they experienced frequent success on course tasks (ALT). Furthermore, if students agreed that *both* First Principles *and* ALT occurred, they were over five times more likely to be rated by their course instructors as high masters of course objectives. When students neither agreed that First Principles occurred nor did they agree that they experienced ALT, they were about 26 times more likely to be rated as low masters of course objectives, compared with agreement that both First Principles and ALT did occur.

The relationship in the Frick et al. (2009) study indicated that First Principles of Instruction are indirectly related to mastery of course objectives. The Spearman correlation between First Principles and mastery was about 0.12, and although statistically significant, it is relatively low. The correlation between First Principles and ALT was much higher ($\rho = 0.58$), and the correlation between ALT and student mastery as determined by their course instructors was 0.36 and highly significant. Thus, it appears that when students agree that their instructors use First Principles of Instruction, it is associated with a greater likelihood of agreeing that they experienced ALT; and in turn, if they agreed they experienced ALT, then they were much more likely to be rated by their instructors as high masters of course objectives, and much less likely to be rated as low masters.

From a theoretical perspective, these patterns make sense. The items on the TALQ Scales used in the present study and also in the Frick et al. (2009) study were derived largely from a synthesis of instructional *theory* on which First Principles of Instruction are based. That theory predicts that when these principles are present, learning is promoted (Merrill, 2002; Merrill, Barclay, & van Schaak, 2008).

The further value of these theoretical principles is that they can be incorporated into a wide range of teaching methods and subject matter. These principles do not prescribe how to teach, nor what to teach. Incorporating First Principles of Instruction into one's teaching may, however, require college instructors to think differently about their subject matter than they are accustomed. Thirty percent of the respondents in this study did *not* agree that First Principles occurred in courses they evaluated, and that was similarly the case in the Frick et al. (2009) study where in 4 of the 12 courses (about 33%), students largely disagreed that First Principles of Instruction occurred. Instead of instruction organized around topics, it may need to be organized on the basis of a sequence of simple-to-complex, whole, real-world tasks or problems (cf. Merrill, 2007). While this can be challenging in redesigning a course, the clear benefit is that such problems or tasks are perceived as more meaningful and relevant by students. When respondents in this study agreed that First Principles occurred (70% of the sample), 9 out of 10 also agreed that they were satisfied with the course, learned a lot, and it was an outstanding instructor/course (see Tables 6, 7, and 8).

Conclusion

We surveyed 193 undergraduate and graduate students from at least 111 different courses at several higher education institutions using a new instrument designed to measure TALQ. Reliabilities ranged from 0.81 to 0.97 for the nine TALQ Scales. Spearman correlations among scales were highly significant, mostly in the 0.60s–0.80s.

Results from APT indicated that students in this study were three to four times more likely to agree that they learned a lot and were satisfied with courses when they also agreed that First Principles of Instruction were used *and* they were frequently engaged successfully (ALT). Students in this study were five times more likely to

agree that they believed they had mastered course objectives when they also agreed that both First Principles and ALT occurred, compared with their absence. Finally, students were almost nine times as likely to rate the course and instructor as outstanding when they also agreed that both First Principles and ALT occurred versus did not occur.

Similar patterns were observed in the Frick et al. (2009) study, and, while fewer classes were observed, most students in each class completed the TALQ instrument. Not only did students self-report their mastery of course objectives, but their instructors independently rated their mastery based on performance in class and on exams, assignments, papers, projects, and other deliverables. Students in that study were about five times more likely to be rated by their instructors as high masters of course objectives, when those students independently reported that they agreed that their instructors incorporated First Principles of Instruction in the course and also agreed that they experienced ALT.

In summary, we believe that the TALQ Scales have considerable promise for use in evaluation of teaching in higher education. These scales are reliable, and scores on these scales are associated with higher student achievement as rated by their instructors. Finally, if instructors receive low evaluations of their teaching on the TALQ Scales on First Principles, these would be areas in which instructors could improve their courses. Such instructors could attempt to build their courses around a series of increasingly complex, authentic tasks (Principle 1); they could make greater efforts to activate student learning (Principle 2); they could model or demonstrate correct task performance more often (Principle 3); they could provide students with more opportunities to try out what they have learned and provide feedback (Principle 4); and they could provide students with more opportunities to integrate what they have learned into their own personal lives (Principle 5). If instructors do increase their use of First Principles, we would expect student ratings on the TALQ Scales to increase, and this in turn should increase the likelihood that more students will master course objectives. Future research studies are needed to empirically determine if this predicted pattern occurs.

References

Abrami, P. (2001). Improving judgments about teaching effectiveness using teacher rating forms. *New Directions for Institutional Research*, *109*, 59–87.

Abrami, P., d'Apollonia, S. & Cohen, P. (1990). Validity of student ratings of instruction: what we know and what we do not. *Journal of Educational Psychology*, *82*(2), 219–231.

American Institutes for Research (2006, January 19). New study of the literacy of college students finds some are graduating with only basic skills. Retrieved January 20, 2007: http://www.air.org/news/documents/Release200601pew.htm.

Arthur, J., Tubré, T., Paul, D. & Edens, P. (2003). Teaching effectiveness: The relationship between reaction and learning evaluation criteria. *Educational Psychology*, *23*(3), 275–285.

Baer, J., Cook, A. & Baldi, S. (2006, January). The literacy of America's college students. American Institutes for Research. Retrieved January 20, 2007: http://www.air.org/news/documents/The%20Literacy%20of%20Americas%20College%20Students_final%20report.pdf.

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.

Brown, B., & Saks, D. (1986). Measuring the effects of instructional time on student learning: Evidence from the Beginning Teacher Evaluation Study. *American Journal of Education*, *94*(4), 480–500.

Clayson, D., Frost, T. & Sheffet, M. (2006). Grades and the student evaluation of instruction: A test of the reciprocity effect. *Academy of Management Learning and Education*, *5*(1), 52–65.

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.

Emery, C., Kramer, T. & Tian, R. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality Assurance in Education*, *11*(1), 37–46.

Feldman, K. (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., et al. (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. (1983). Non-metric temporal path analysis: An alternative to the linear models approach for verification of stochastic educational relations. Bloomington, IN. Retrieved, March 4, 2007: http://www.indiana.edu/~tedfrick/ntpa/.

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., An, J. & Koh, J. (2006). Patterns in Education: Linking Theory to Practice. In M. Simonson (Ed.), Proceedings of the Association for Educational Communication and Technology, Dallas, TX. Retrieved March 4, 2007: http://education.indiana.edu/~frick/aect2006/patterns.pdf.

Frick, T., Chadha, R., Watson, C. & Zlatkovska, E. (2009, under review). Improving course evaluations to improve instruction and complex learning in higher education. Submitted to *Educational Technology Research & Development*.

Kirk, R. (1995). Experimental design: Procedures for the behavioral sciences (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kirkpatrick, D. (1994). *Evaluating training programs: The four levels*. San Francisco, CA: Berrett-Koehler.

Koon, J., & Murray, H. (1995). Using multiple outcomes to validate student ratings of overall teacher effectiveness. *The Journal of Higher Education*, *66*(1), 61–81.

Kuh, G., Kinzie, J., Buckley, J., Bridges, B., & Hayek, J. (2006, July). What matters to student success: A review of the literature (Executive summary). Commissioned report for the National Symposium on Postsecondary Student Success. Retrieved January 20, 2007: http://nces.ed.gov/npec/pdf/Kuh_Team_ExecSumm.pdf

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

Marsh, H. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, *76*(5), 707–754.

Merrill, M. D. (2002). First Principles of Instruction. *Education Technology Research and Development*, *50*(3), 43–59.

Merrill, M. D. (2007). A task-centered instructional strategy. *Journal of Research on Technology in Education*, *40*(<iss>1</iss>), 33–50.

Merrill, M. D., Barclay, M. & van Schaak, A. (2008). Prescriptive principles for instructional design. In J. M. Spector, M. D. Merrill, J. van Merriënboer & M. F. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 173–184). New York: Lawrence Erlbaum Associates.

Renaud, R., & Murray, H. (2004). Factorial validity of student ratings of instruction. *Research in Higher Education*, *46*(8), 929–953.

Squires, D., Huitt, W. & Segars, J. (1983). *Effective schools and classrooms: A research-based perspective*. Alexandria, VA: Association for Supervision and Curriculum Development.