Running head: THEORY-BASED COURSE EVALUATION

Theory-Based Course Evaluation:

Implications for Improving Student Success

in Postsecondary Education

Theodore W. Frick,

Rajat Chadha,

Carol Watson,

Ying Wang, and

Pamela Green

Department of Instructional Systems Technology

School of Education

Indiana University Bloomington

Submitted to the American Educational Research Association Conference, 2008

March 1, 2008

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. TALQ, a new course evaluation instrument, was developed from validated instructional theory that predicts student success. A survey of 193 students in 111 different courses at multiple institutions was conducted using TALQ. Results indicated strong associations ($p < 0.0286$) 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.

Problem

This study began because the first author served on a university committee which 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 .43; the average overall course rating and student achievement was .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, p. 193), 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 which 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 to 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 explain a relatively small proportion of variance in student learning achievement (Emery, Kramer & Tian, 2003). In a more recent example, Arthur, Tubré, Paul & 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 two- and four-year public universities and colleges. They used the same standardized assessment instrument as that in the National Assessment of Adult Literacy (2003). The literacy assessments were supervised by a test administrator on each campus.

The Baer *et al.* report provides some sobering findings. They reported percentages of students from 2-year vs. 4-year institutions, respectively, who were *proficient* in prose literacy as 23% and 38%, in document literacy as 23% and 40%, and in quantitative literacy as 18% and 34%. 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.* made several recommendations. One important recommendation was to "… *focus assessment and accountability efforts on what matters to student success*" (p. 4, italics added).

Revisiting the Content of Course Evaluations with a Focus on Student Success

Results from these recent studies provide impetus for reexamining the kinds of items used on typical course evaluations in higher education. Can we develop better scales to measure factors that are empirically known to be associated with higher levels of achievement? If so, then perhaps we can use new course evaluation ratings with greater validity and utility than those traditionally used. 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.

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, 1990; 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. Can we measure student ALT with a course evaluation instrument?

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. Merrill 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). Can we measure first principles of instruction with a course evaluation instrument?

Levels of evaluation of training. Finally, we considered levels of evaluation of training effectiveness that have been used for more than five decades in non-formal 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 learners belong.

[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.

With respect to Level 2, student learning achievement, we wondered if we could get good estimates from students themselves. 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 multi-section courses. We asked students about their learning achievement in several different ways: grades received or expected, mastery of course objectives, and about how much they believed they had learned.

Method

A survey instrument was constructed that contained items intended to measure scales for student ratings of self-reported academic learning time, satisfaction with the course, learning achievement, 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. Each set initially contained five items intended to measure the respective construct (scale). 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://domain.blinded.during.review.edu/xyz .

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 completed the survey who had nearly or recently completed a course between May, 2006 and June, 2007. One hundred ninety-three valid cases remained, after elimination of those containing no data, several test cases to ensure that data collection was working as intended via the Web survey, those which had 6 or 7 error flags, and those which were multiple submissions from the same respondent.

There were 7 pairs of items that were stated oppositely and individually scattered randomly throughout the instrument (e.g., items 5 and 35). This was done intentionally, in order to identify respondents who were not reading the items carefully and rating them all similarly. The PHP program which processed the survey data set an error flag whenever a respondent agreed or strongly agreed with a pair, or disagreed or strongly disagreed with that pair. In analysis of those respondents' data sets, we observed that they very often checked the same Likert response to all 45 items (e.g., agree) and completed the survey in just a few minutes (the PHP software also determined how long the respondent took to complete the survey). We did not eliminate cases with less than 6 error flags, since we expected the reliability analyses to identify poor items.

The PHP software also captured the IP address of the computer on which the survey was completed. If two data sets were submitted in succession with the same IP number on the same day and time, only a few seconds apart, we assumed that the respondent clicked the final submit button twice (and this was confirmed by observing identical or nearly identical data sets for those two cases). Where this occurred, we removed the first submission. Of the 193 respondents, 162 IP numbers were unique. We did not remove cases with the same IP numbers which were clearly different data sets and not submitted at the same time, since these likely came from students working in a university computer lab at the same computer at different times. Roughly 33 percent of IP addresses were from .edu domains, 40 percent from .com, and 25 percent from .net domains. Of the .edu domains, three campuses from the authors' institution comprised approximately one-third of the respondents who completed the survey on campus computers. This means that about two-thirds of the respondents most likely completed the survey via their local Internet Service Providers. There was a very wide range of ISP's, with no one service provider dominant.

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). We list them below, because these qualitative data reflect a very wide range of course topics and disciplines, which is important with respect to interpretation of results from this study and its generalizability: addictions counseling; advanced educational research; algebra; anthropology;

applied research; assessment strategies in education; behavioral pharmacology; bilingual education literacy; biology lab; business finance; business and society; business law honors; cell biology; cognition and technology; cognitive theories; combinatronics; communication in electronic environments; community based anesthesiology; comparative education; computers in education; creative writing; critical care medicine; dance; database - Microsoft Access; database design; database management; death and dying; developing websites for learning; differentiation for all learners; dissertation proposal preparation; dynamic systems theory in cognitive science; curriculum and methods for students with severe disabilities; educational psychology; educational research methodology; English; English and literature; English composition; English essay writing; English writing; family and marriage; family development over life; finite math; foundations of doctoral study; foundations of graduate study; fundamentals of math; general and systemic pathology; graduate seminar; graphic design using Microsoft Publisher; gross human anatomy; history of epidemics in the new world; human biology; biology; independent study; information, people and technology; inside of business world; instructional design basics; instrumental and choral conducting; intermediate statistics; internal medicine, VA hospital; introduction to public and community health; introduction to American politics; introduction to business; introduction to business administration; introduction to college writing; introduction to physical education; introduction to psychology II; introduction to statistics; learning and employability; learning to work in groups; managing and empowering students; mathematical statistics; mechanism of human disease; medical biochemistry; medical genetics; medical pathology; medical physiology; methods of action research; microcomputer and computer business graphic applications; microeconomics; mixed methods in educational research; moral controversies in American politics; music theory; needs and task analysis; online library research; organizational behavior; organizational management; pathology; pathophysiology; pediatrics; pharmacy technology; plagiarism; pocket PC applications; professional writing; psychology; psychology as a discipline or profession; seminar on educational technology; social psychology; social studies for elementary teachers; sociology - family violence; spectroscopy; statistics; teaching and learning in higher education; teaching language arts in elementary and middle grades; theory of knowledge; topics on diversity and social work; web course development; work and communication with different people in organization; writing; and educational assessment and measurement.

While courses in business (34), medicine (23), education (18), English (18), and computers and technology (12) were mentioned more frequently than others, it can be seen that 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 either 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 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 percent in each group).

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

Insert Table 1 here

Course grades. Table 1 also displays responses of students with respect to their course grade. Almost 93 percent reported that they received or expected to receive an A or 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 percent reported themselves to be masters. The large majority considered themselves to be partial masters of course objectives, while 16 percent 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 to 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 ANOVA's 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 ($\rho = 0.306$, $p < 0.0005$, $n = 191$). 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 their reliabilites are listed in Tables 2.1 to 2.10. While not reported in these tables, a frequency analysis of Likert ratings indicated that for most of the 35 positively

[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

stated items respondents were about twice as likely to agree or strongly agree with the items as not. The same pattern obtained in reverse for most of the 10 negatively worded items.

To determine the reliability of each scale, all 5 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 Table 2.1 that internal consistency of each scale was generally quite high.

Insert Tables 2.1 and 2.2 here

Combined First Principles scale (Merrill 1 to 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 scores into the reliability analysis, treating each principle score as an item score itself. The resulting Cronbach 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 *ρ* is computed by first converting the scale score for each case to a rank, and then a Pearson Product Moment Coefficient is calculated on the ranks.

The correlations are presented in Tables 3 and 4. 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 $p < 0.0286$.

Insert Tables 3 and 4 here

First Principles of Instruction considered individually. It can be seen in Table 3 that First Principles are highly correlated with each other, with all correlations significant at $p < 0.0005$, with ρ ranging from 0.722 to 0.819. This should not be surprising, since the internal consistency α was 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 4 are very strong as a group. Except for student mastery, the Spearman correlations ranged from 0.46 to 0.89, with most in the 0.60's to 0.80's. 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 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 (APT).

Pattern Analysis (APT)

While there were numerous highly significant bivariate relationships which explained typically between 40 and 80 percent of the variance in ranks, specific patterns that show temporal relations among 3 or more variables are not shown in Tables 3 and 4. 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, ANOVA, etc. – 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, An & Koh, 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.[6] A scale was recoded as “Yes” if the scale

[6] 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.

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. This method is nearly equivalent to choosing the modal rank for each scale for each case, but SPSS 14 had no provision for such computation. A sample of test cases, where modal ranks were recoded by hand, indicated a nearly perfect correlation between this method and codes generated by the above computational procedure using scale means and a 3.5 cut-off.

If *ALT* **and** *First Principles,* **then** *Learned a Lot*. In Table 5 results are presented for the APT pattern: 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.

Insert Table 5 here

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 4 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.

Insert Table 6 here

If *ALT* **and** *First Principles,* **then** *Learner Satisfaction*. In Table 6, 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 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 to 1 (0.96/0.29).

Insert Table 7 here

If *ALT* **and** *First Principles,* **then** *Outstanding Instructor/Course*. 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 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 to 1 that an Instructor/Course is viewed as outstanding by students when ALT and First Principles are both present versus both absent, according to student ratings.

Insert Table 8 here

If *ALT* **and** *First Principles,* **then** *Mastery*. 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 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 5 times more likely to agree that they mastered course objectives when they agreed vs. did not agree that both ALT and First Principles occurred when they took the course.

Factor Analysis: Teaching and Learning Quality – A Single Trait?

Spearman correlations among the scales measured in this study were generally very high. Are these scales measuring the same overall construct, perhaps something that might be called 'Teaching and Learning Quality?' To investigate this possibility, we conducted a factor analysis of the scales reported in Table 4. We used the image analysis extraction method. The image method of factor extraction, "distributes among factors the variance of an observed variable that is reflected by the other variables – and provides a mathematically unique solution" (Tabachnick & Fidell, 2001, p. 612). The net effect of this approach is to minimize the impact of outliers (e.g., see Veldman, Kaufmann, Agard & Semmel (1985), p. 55).

Results of factor analysis are presented in Table 9. A single factor was extracted which accounted for nearly 70 percent of the variance. Remaining factors had eigenvalues less than one. The factor loadings in Table 9 ranged from 0.94 for learner Satisfaction to 0.35 for Mastery Level.

Insert Table 9 here

These results indicate that the scales used in this study may be measuring a unidimensional trait.

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).

Similar to the notion of *transfer* in qualitative methodology, college instructors can test APT findings in their own contexts. Instructors can test APT predictions by using the scales from this study to evaluate a course currently taught, and by considering this evaluation as a baseline measure. Instructors should also measure student achievement in the course by objective means. Then instructors can try modifying their courses and actions to increase ALT and use First Principles throughout the course. Then repeat use of the scales from this study to evaluate the new version of the course. Do the ratings improve for this redesigned course? Do student achievement scores also increase, as measured by objective course exams or performance assessments?

Does increased ALT and use of First Principles *cause* increased student learning achievement and satisfaction? It would be hard to say from just one experiment such as the one suggested above. However, if this sort of pattern repeatedly occurs for many instructors and their courses, then this would further increase confidence in the prediction.

Mastery of learning objectives. As noted earlier, less than 1 out of 4 students considered themselves masters of course objectives, even though 93 percent received A's and B's for their course grades. It appears that students could be learning more in their courses.

Student self-reports of mastery are consistent with the study done by Baer, Cook and Baldi (2006) which reported on the accomplishments of a nationwide sample of college students in 2- and 4-year institutions. Over 1,800 students at 80 randomly selected colleges and universities were independently tested (i.e., not by their instructors) on practical skills in prose literacy, document literacy, and quantitative literacy. More than 75% of students at 2-year institutions performed lower than proficiency level, and more than 50% at 4-year institutions likewise scored below proficiency level. These are practical life skills that many college students have not mastered.

Even though we were not able to randomly sample students as did Baer, Cook and Baldi, students in our study rated a wide range of courses and topics (at least 111 unique courses in business, health sciences, education, and the liberal arts). The consistency between these two studies supports the generalizability of findings from our study.

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 which 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. First Principles were correlated highly with student self-reports of learning a lot ($\rho = 0.83$).

Finally, 25 items on this survey 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. 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. They may, however, require college instructors to think differently about their subject matter than they are accustomed (30 percent of the respondents in this study did *not* agree that First Principles occurred in courses they evaluated). Instead of instruction organized around topics, instruction may need to be organized on the basis of a sequence of simple to complex, whole, real-world tasks or problems (cf. Merrill, 2007, in press). 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 percent of the sample), 9 out of 10 also agreed that they were satisfied with the course, learned a lot, and that it was an outstanding instructor/course (see Tables 5 - 7).

Will students learn more and achieve more if instructors incorporate First Principles into their courses? We cannot conclude that incorporation of First Principles *causes* students to learn. In general, we cannot *logically* conclude that instruction is a necessary condition for learning to occur, since learning can clearly occur in the absence of any instruction or instructional program.

Nor is instruction a sufficient condition for learning to occur. Learners are not like light bulbs, where someone can flip a switch that causes learning to occur. When learners engage in tasks and activities, this is ultimately a matter of choice that learners make. Learners are intentional systems (Thompson, 2006). When they do make that choice, and their engagement is successful, there is clear evidence in the Academic Learning Time (ALT) literature that this is associated with higher levels of student achievement. Instructors can try to create conditions and activities that will increase the likelihood that students will make that choice to engage and try to learn. And when students do so, instructors can provide feedback to reduce errors and increase successful performance (Principle 4).

While temporal patterns do not imply causation, this does not mean we cannot act until causation is proven. In the 1960s, the U.S. Surgeon General mandated that warnings be put on cigarette packs that smoking may cause lung cancer. Physicians had observed that smokers were more likely to get lung cancer later in their lives if they smoked. Cigarette makers argued for years that smoking does not cause lung cancer. Nonetheless, heavy cigarette smokers are 5-10 times more likely to have lung cancer later in their lives (Kumar, Abbas & Fausto, 2005), and if they quit smoking the likelihood decreases. While causal conclusions cannot be made in the absence of controlled experiments, nonetheless one can make practical decisions based on such epidemiological evidence. We can do likewise with APT results, particularly when they are consistent with theoretical predictions.

In the Special Theory of Relativity, light was predicted to bend when passing objects of great mass which results in the curvature of space. The theory led physicists to measure the deflection of light passing near our sun from the position of stars during a solar eclipse and compare the positions of those stars when the sun was not near the line of sight. Einstein's theory predicted the amount of deflection that would be expected, so that the observed results could be compared (Einstein, 1961, p. 129).

Physicists probably would not have thought to measure such deflections during a solar eclipse had Einstein's theory not implied such a prediction. This illustrates the value of theory (cf. Thompson, 2006). There was no experiment, no randomized trials – just a theory and some evidence to support it.

Based on several theories of instruction, Merrill (2002) claimed that learning will be promoted when First Principles of Instruction are utilized. That prediction was supported in our study by the reports by students on their learning experiences in college courses. We would not have thought to put such items on our survey instrument without such a prediction. We would not have thought to make the APT queries that we did in Tables 5 to 8.

Conclusion

We surveyed 193 undergraduate and graduate students from at least 111 different courses at several institutions using a new instrument designed to measure teaching and learning quality (TALQ). Reliabilities ranged from 0.81 to 0.97 for the nine TALQ scales. Spearman correlations among scales were highly significant, mostly in the 0.60's to 0.80's. Factor analysis indicated that the TALQ scales may be measuring a single dimension of teaching and learning quality in postsecondary institutions as students perceive it.

Results from analysis of patterns in time (APT) indicated that students in this study were 3-4 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 5 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 9 times as likely to rate the course and instructor as outstanding when they also agreed that both First Principles and ALT occurred vs. did not occur.

As the saying goes, "It takes two to tango." Even if instructors provide authentic problems to solve, activate student learning, and demonstrate what is to be learned, students themselves must also try to learn. Students must engage in solving those problems so that instructors can coach them and give guidance and feedback as needed. Instructors can encourage students to integrate what they have learned into their own lives, but it is the students who must do that integration.

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.

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.

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 .

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.

Einstein, A. (1961). Relativity: The special and general theory (translated by R. W. Lawson). NY: Crown Publishers.

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., & 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. (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. and 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 .

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. & 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

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

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

Kumar, V., Abbas, A., Fausto, N., (2005). *Robbins and Cotran pathologic basis of disease* (7th edition). Philadelphia: Elsevier/Saunders.

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 & Development, 50*(3), 43-59.

Merrill, M. D. (2007, in press). A task-centered instructional strategy. *Journal of Research on Technology in Education*.

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.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics (4th ed.).* Boston: Allyn and Bacon.

Thompson, K. R. (2006). Axiomatic theories of intentional systems. *Scientific Inquiry Journal, 7*(1), 13-24. Retrieved July 4, 2007: http://www.iigss.net/Scientific-Inquiry/THOMPSON-2.pdf .

Veldman, D., Kaufman, M., Agard, J. & Semmel, M. (1985). Data reduction and analysis. In Kaufman, M., Agard, J., and Semmel, M. (Eds.), *Mainstreaming: Learners and their environments.* Cambridge, MA: Brookline Books.

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

Table 2.1 Nine TALQ scales

1. Academic Learning Time Scale (**α = 0.81**)

Item No.	*Item Stem*[7]
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.

2. Learning Scale (α = 0.95)

Item No.	*Item Stem*
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.

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

32- I did not learn much as a result of taking this course.

3. Global rating items selected from the standard university form (α = 0.97)

Item No.	*Item Stem*
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 (α = 0.87)

Item No.	*Item Stem*
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 (α = 0.90)

Item No.	*Item Stem*
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 (α = 0.89)

Item No.	*Item Stem*
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 (α = 0.82)

Item No.	*Item Stem*
7	My instructor detected and corrected errors I was making when solving problems, doing learning tasks or completing assignments.
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 (α = 0.87)

Item No.	*Item Stem*
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 (α = 0.94)

Item No.	*Item Stem*
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 2.2. Combined First Principles Scale (α = 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

Table 3. Spearman's ρ correlations for First Principles of Instruction scales

		Authentic Problems	Activation	Demonstration	Application	Integration
Authentic Problems Scale	ρ	1.000				
	N	192				
Activation Scale	ρ	.790**	1.000			
	N	192	193			
Demonstration Scale	ρ	.803**	.792**	1.000		
	N	189	190	190		
Application Scale	ρ	.724**	.763**	.794**	1.000	
	N	186	186	184	186	
Integration Scale	ρ	.819**	.818**	.770**	.722**	1.000
	N	192	193	190	186	193

** Correlation is significant ($p < 0.0005$, 2-tailed).

Table 4. Spearman's *ρ* correlations among scales

		First Principles	ALT	Learning	Satisfaction	Global Rating	Class Rating	Mastery
Combined First Principles	*ρ*	1.000						
	N	193						
ALT	*ρ*	.670**	1.000					
	N	192	192					
Learning	*ρ*	.833**	.747**	1.000				
	N	193	192	193				
Satisfaction	*ρ*	.850**	.683**	.856**	1.000			
	N	192	191	192	192			
Global Rating	*ρ*	.890**	.605**	.811**	.903**	1.000		
	N	193	192	193	192	193		
Class Rating	*ρ*	.694**	.464**	.649**	.753**	.773**	1.000	
	N	191	190	191	190	191	191	
Mastery of Objectives	*ρ*	.344**	.359**	.334**	.317**	.341**	.306**	1.000
	N	192	191	192	191	192	191	192

** Correlation is significant (p < 0.0005, 2-tailed).

Table 5. APT Results for the Pattern: If *ALT* and *First Principles,* then *Learning*

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

Table 6. APT Results 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

Table 7. APT Results for the Pattern: If *ALT* and *First Principles,* then *Outstanding Instructor/Course (BEST)*

	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 8. APT Results 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

Table 9. Factor Matrix for Main Scales

	Factor
	1
Satisfaction Scale	.940
Global Rating Scale	.936
Combined First Principles	.908
Learning Scale	.869
Class Rating	.820
ALT Scale	.643
Mastery Level	.346