Theory-Based Course Evaluation:
Nine Scales for Measuring Teaching and Learning Quality

Theodore W. Frick,
Rajat Chadha,
Carol Watson,
Ying Wang, and
Pamela Green

Department of Instructional Systems Technology
School of Education
Indiana University Bloomington

March 12, 2007

Abstract

Traditional course evaluations in higher education contain few items which are strongly related to student achievement. The *best* predictors of student achievement are typically global items that correlate only *moderately* with student achievement. Can better items be used that are based on instructional theory and research? We developed a survey containing nine *a priori* scales and received 140 responses from students in 89 undergraduate and graduate courses at multiple institutions. Data analysis indicated strong correlations between academic learning time, student achievement, first principles of instruction, student satisfaction, mastery of course objectives, and global course ratings. Most importantly, these scales measure principles through which instructors can improve their classes: provide authentic problems for students to solve; activate prior learning; demonstrate what is to be learned; provide repeated opportunities for students to successfully solve problems with coaching and feedback; and help students integrate what they have learned into their personal lives.

Problem

This study began because the first author served on a 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 a philosophy of teaching statement, letters from students and colleagues, samples of student work, course syllabi, publications related to teaching, and course evaluation results during the past year. These were customary forms of evidence that have been used in the past, and similar to those used for evaluation of teaching for promotion and tenure. This experience nonetheless raised the question: What empirical evidence is there that any of these indicators are associated with student learning achievement?

Thus, the first author began to look at 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 is little doubt that ratings by students of instructors and courses in higher education have demonstrable validity.

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 & 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

Perhaps the issue of course evaluation should be further examined in light of what appears to be unsatisfactory levels of student achievement at colleges. 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 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; Fisher, et al., 1978; 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, Merrill (2002) synthesized instructional design 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* (students solve a series of increasingly complex real-world problems); 2) *Activation* (students 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 themselves with scaffolding and feedback from instructors or peers); and 5) *Integration* (students are able 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.

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.

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. See Tables 2.1 to 2.9 for the nine *a priori* item sets. Each set contained five items intended to measure the respective construct (scale). The reader should note that a minimum of two items is needed to determine internal consistency of a scale

[1] It should be also noted that Kirkpatrick's Level 3 is very similar to Merrill's Principle 5 (integration).

(Cronbach's α) on a single instrument administration. Past experience in large-scale research studies by the first author indicated that more than five items per scale were unlikely to improve internal consistency.

A paper version of the instrument was then reviewed and wording of items considered to be confusing or ambiguous was modified. In particular, much of the discussion involved items about real-world problems or authentic tasks. Instructors were concerned about the perceived meaning of these terms by students, and thus an explanatory note was added to each page of the survey to provide a definition of 'authentic problems' or 'authentic tasks' as "meaningful learning activities that are clearly relevant to you at this time, and which may be useful to you in the future (e.g., in your chosen profession or field of work, in your life, etc.)."

The instrument, now referred to as the *Teaching and Learning Quality Scales* (abbreviated as the *TALQ Scales*), was then converted to a Web survey, which can be viewed online at: http://education.indiana.edu/~edsurvey/evaluate/ . A study information sheet was required by the University's institutional research board (IRB). Before students could view the survey itself, they were informed that:

> The purpose of this study is to examine relationships among items on course evaluation forms, instructional practices, and student academic learning time… (y)ou will complete a Web-based survey about one of your current or recent classes, your participation in course activities, and your sense of accomplishment in that course. The data will be compiled by the researchers. No individual student data will be seen by the instructor of your course.

The reader should note that no mention was made of Merrill's first principles of instruction or Kirkpatrick's levels of evaluation. Furthermore, student ratings were not shared with their instructors and hence could not affect their grade in the course.

The Web survey was written in PHP and HTML and published on a university server. Items were randomly ordered for the main scales (listed in Tables 2.1 through 2.9). Furthermore, the PHP software was designed to detect inconsistent responses to scale items in order to detect cases in which someone may have clicked the same response to all items on a scale, when some of those items were negatively worded. These were "error flags" that were set when this condition was detected. In addition to responses, the IP number of the respondent's computer, the current date and time, and the number of seconds it took to complete the survey were stored in a data file on the server. On the first page of the survey, information was requested about the title or subject matter of the course, instructor name(s), student overall perception of the quality of the course, student gender, expected or received grade, mastery level, class standing, and whether the course was face-to-face, online or blended.

Multiple requests for participants were sent via e-mail. By January 25, 2007, a total of 156 responses had been recorded in the survey data file.

By examining the error flags discussed above, those cases with six error flags were identified. Three of these were test cases submitted by the first author over the nine-month data collection period to insure that the data collection system was working (and should have been deleted accordingly), but the remainder were cases in which there were no responses to items on the nine *a priori* scales. These cases were deleted, since there was no data we could analyze from them. We also checked to make sure the same individual did not submit the same data more than once, e.g., by clicking the final "submit" button twice. We did *not* eliminate any cases with error flags that actually contained responses to any of the nine scales, since we wanted to evaluate the instrument itself for internal consistency. We did not eliminate any cases which students listed one of the researchers as the instructor, but these constituted only six cases out of the original 156. Before beginning data analysis, a total of 16 cases were eliminated based on problems identified in the data file with those cases. The resulting data set for analysis was 140 cases.

The data were downloaded from the server and imported into SPSS 14 for analysis. Prior to reliability and subsequent analyses, responses to items with negative wording were reverse-scored and new variables were created in SPSS.

Results

Since participation in the survey was voluntary, we collected demographic data in order to aid in the interpretation of results and to get an idea of the representativeness of the obtained sample of 140 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 (text) to the question about the course title or content. A total of 89 different subject areas were mentioned by 130 respondents that included: educational technology, diversity and social work, medical physiology, history of world epidemics, educational leadership, spectroscopy, medical biochemistry, genetics, educational assessment, pathology (human disease), independent study, cell biology, critical care medicine, pediatrics, dance, internal medicine, social studies education, bilingual education literacy, human cognition and theories, human anatomy, anesthesiology, business administration, organizational behavior/management, introduction to business, professional writing, finite mathematics, mathematical statistics, educational research, introduction to psychology, American politics, business law, epistemology, teaching and learning in higher education, web development, curriculum and instruction, business finance, intermediate statistics, pharmacology, instructional design, music theory, PC applications, systems theory, business graphics, addictions counseling, managing students, comparative education, business and society, writing, database management, biology, sociology, pharmacy technology, graphics design, educational measurement, algebra, doctoral study, teaching language arts, research methodology, anthropology, social psychology, graduate seminar, fundamentals of mathematics, physical education, instrumental/choral conducting in music, educational psychology, and biology laboratory.

While courses in business (33), medicine (22), education (18), and computers and technology (13) were mentioned more frequently than others, it can be seen that a wide range of subjects were represented in the courses taken by respondents.

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 were listed more than once by different respondents, the 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 93 females and 43 males responded to the survey (4 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 associations between gender and academic learning time, learning achievement, student satisfaction, first principles, grades, and course ratings. Additionally, there were no significant associations between gender and other demographics, except for two weak relationships 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 - 16 percent in each group).

Course settings. Nearly 70 percent of courses evaluated were face-to-face, and about one-fourth were online or distance courses.

Insert Table 1 here

Course grades. Table 1 also displays responses of students with respect to the grade they either expected to receive in the course they were evaluating or which they did receive. Approximately two-thirds got A's and about 21% B's. It is unclear from these data alone whether this may be evidence of grade inflation, or an indication that those respondents who were higher achieving students were more likely to respond to the survey than lower achieving students – since participation was voluntary.

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 only 25 percent of respondents reported that they had mastered course objectives, even though 87 percent received A's or B's (see above). About 62 percent of students considered themselves as "partial masters" of course objectives and 12 percent reported themselves as "nonmasters".

Relationships among Variables

With the exception of gender, variables course grades, mastery, gender, class standing, and overall course ratings are ordinal; thus, Spearman's *rho* (ρ) was used to measure their association. Spearman's ρ is a correlation of ranks, in contrast to Pearson product moment correlation coefficients. With ρ there is no assumption that measures of variables are normally distributed nor that the intervals between measurement units are equal, as assumed in Pearson correlations. Computationally, the values for each variable are converted to ranks first, and then a Pearson product moment coefficient is computed on the ranks in order to determine ρ.

In this study, we choose our *a priori* Type I error rate as $\alpha = 0.0005$ for determining statistical significance. The reason for this is that our sample size was fairly large ($n = 140$ cases) and we sought to minimize the probability of concluding statistical significance as an artifact of making many comparisons. Kirk (1995, p. 120) emphasizes that the actual Type I error rate is equal to $1 – (1 – \alpha)^C$. We expected to conduct about 50 statistical tests. We ultimately conducted a total of 58 statistical tests. The overall Type I error rate for this study is $1 – (1 – 0.0005)^{58} = 0.0286$. Thus, the Type I error rate for our study was less than 0.05 or about 0.03.

Finally, statistical significance is less important than proportional reduction of error in making predictions. If the sample size is large enough, small differences can be statistically significant but have little practical significance. Proportional reduction of error in correlational studies is indicated by the square of the correlation coefficient, or proportion of variance that is predictable in one variable by another (Ferguson, 1971).

Gender. Since gender is a nominal level variable, chi square tests were performed. Gender (0 = male, 1 = 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 ($\chi^2 = 6.27$, $df = 2$, $p = 0.043$, $n = 136$) 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. There was a weak relationship between gender and class standing. An ANOVA was performed, resulting in $F = 21.94$, df=1,134, $p = 0.004$. The average male was a senior (mean = 3.81) and the average female a junior (mean = 2.96). A chi-square analysis indicated that a few more males were graduate students than expected, and a few more females were freshmen than expected. However, the chi-square was not significant ($p = 0.135$), even though the linear association approached significance in the ANOVA.

Student mastery level. There was a significant association between class rating and mastery of course objectives ($\rho = 0.319$, $p < 0.0005$, $n = 138$). Students who considered themselves masters of course objectives were more likely to rate the course as "great" ($\rho^2 = 0.101$, or about a 10 percent reduction in error in predicting mastery level based on knowledge of course rating rank, or vice-versa).

There was also a significant correlation between student reports of mastery level and course grades ($\rho = 0.373$, $p < 0.0005$, $n = 129$). This represents about a 14 percent reduction in error in predicting mastery based on knowledge of grade received.

[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, 0 = other

Other variables. None of the remaining associations among these variables was statistically significant at $p < 0.0005$. It is noteworthy that students' expected or received course grades were very weakly associated with their ranks of overall course quality ($\rho = 0.192$, $p = 0.030$, $n = 128$, $\rho^2 = 0.037$).

Scale Reliabilities

Tables 2.1 through 2.9 provide descriptive statistics on the 45 items from the survey, grouped according to *a priori* scales. It can be seen that, in general, for most of the 35 positively stated items that 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. The choices for each item were a standard Likert rating scale, ranging from "strongly disagree" (coded as 1) to "strongly agree" (coded as 5). In addition, each item had a "not applicable" choice, which was treated as missing data in all analyses.

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. Maximum α is 1.00, and an α of 0 means no reliability (and even negative values of α can be obtained when items are negatively correlated). Items that were negatively worded (-) had their Likert scores reversed by computing new variables in which 5 became 1, 4 became 2, 3 stayed the same, 2 became 4 and 1 became 5.

SPSS provides an option to compute the α coefficient if an item is removed from the scale. Since one of our goals was to reduce the number of items while maximizing scale reliability, we removed an item from the scale if α would increase by such removal. Items were removed one at a time until no item could be removed without decreasing the α coefficient. Item stems with strikethroughs were those removed under each scale according to this procedure. The reliability reported for each scale is that with the remaining items on that scale. This is a standard procedure for scale construction.

It should be noted that *individual items themselves have no reliability* when determining internal consistency reliabilities. At least two items are necessary to form a scale. It should be noted that factor analysis was not considered appropriate at this point, since these scales were formed *a priori* based on what we were trying to measure: academic learning time, student achievement, global course rating, authentic problems, activation, demonstration, application, integration and learner satisfaction.

Our goal was to form a single scale score for each reliable scale before further analysis of relationships among variables measured in the study. Adequate reliability, in the classical sense of reliability of measures, is a necessary but not a sufficient condition in order to find significant relationships among variables (Frick & Semmel, 1978).

Insert Table 2.1 here

Academic learning time (ALT). It can be seen in Table 2.1 that two items were eliminated from the ALT scale resulting in α = 0.85. Both of these items were negatively worded (1 and 29). This ALT scale is measuring student agreement with being 'frequently engaged in successfully completing learning tasks or solving problems in the course.' The items on this scale are consistent with the definition of academic learning time as reported in the literature (e.g., Berliner, 1991; Brown & Saks, 1986; Fisher et al., 1978).

Insert Table 2.2 here

Learning achievement scale (Kirkpatrick, Level 2). None of the 5 items on this scale could be removed without decreasing the α reliability which was 0.97. This scale is measuring student agreement with 'learning a lot in the course, compared with when I began.' While this is a self-report of achievement, Cohen (1981) and others (e.g., Feldman, 1989; Kulik, 2001) have reported that student self-assessment of their own learning achievement is often related positively with external measures such as exam scores. In other words, students have a fairly good idea most of the time about whether they are learning something or not.

Insert Table 2.3 here

Global course evaluation scale (from BEST). Two items were eliminated from this scale, resulting in an α of 0.92. This scale measures student agreement with the ‘quality of the course and instructor as being outstanding.’ This scale is consistent with the global ratings that Cohen (1981) and others (e.g., Kulik, 2001) have reported in meta-analyses as being most highly related to measures of student achievement in multi-section courses with common examinations (average correlations of .43 and .47 in Cohen’s meta-analysis).

Insert Table 2.4 here

Authentic problems scale (Merrill principle 1). One item was eliminated from this scale, resulting in an α of 0.81. This scale measures student agreement with ‘solving a series of increasingly complex real-world problems.’

Insert Table 2.5 here

Activation scale (Merrill principle 2). No items could be removed from this scale without decreasing the reliability. Cronbach’s α was 0.91 for this scale, which measures student agreement with ‘my instructor helped me to link past learning or experience with what is to be newly learned.’

Insert Table 2.6 here

Demonstration scale (Merrill principle 3). One item was deleted from this scale resulting in an α of 0.88. This scale measures student agreement with ‘exposure to differentiated examples of what they are expected to learn or do.’

Insert Table 2.7 here

Application scale (Merrill principle 4). One negatively worded item was removed from this scale. Cronbach’s coefficient α was 0.74. This scale measures student agreement with ‘solving problems themselves with scaffolding and feedback from instructors or peers.’

Insert Table 2.8 here

Integration scale (Merrill principle 5). All five items were needed for this scale, the reliability of which was 0.81. The integration scale measures student agreement with ‘being able to use what I have learned in my own personal life.’

Insert Table 2.9 here

Learner Satisfaction scale (Kirkpatrick, Level 1). Two items were removed from this scale, resulting in an α of 0.94. This scale measures student agreement with ‘being satisfied with this course.’

Combined First Principles scale (Merrill 1 to 5). We were further interested in the extent to which the First Principles themselves formed a reliable scale. To do this, we first formed a scale score for each First Principle by adding its respective item scores and dividing by the number of items on this scale. Thus, the scale score represented the average Likert rating for that scale 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 scale scores. Scores were created for remaining scales as explained above for the first principles scales. For example, for the Learner Satisfaction scale for each case we added the reversed scores for items 6 and 20 and the obtained score for item 45, and then divided this sum by 3 (since there were 3 items on this scale). Thus, for each case, each of the nine scales resulted in an average Likert rating. For example, if the scores were 4, 4 and 5 for a case on the Satisfaction Scale, the average is 4.33, and when rounded is 4. Thus, for this case we can say that he or she 'agreed' with 'being satisfied with this course.'

Finally, it is very important to note that if there were missing data on any of the items for a scale for a given case, then SPSS did not compute the average but put a missing value as the score for that scale score for that case. This did reduce the total number of cases with valid scores on each scale, as will be seen in subsequent analyses. We do not believe that this affected our results and conclusions, since there were plenty of cases left.

Correlational Analyses

We next investigated the relationships among the scales themselves. Since Likert scale scores are ordinal level data in the first place, Spearman's ρ was used as a measure of association, for reasons explained earlier. In other words, we assume that 'strongly agree' is better than 'agree' but we do not assume that the *interval* between 'agree' and 'strongly agree' is equal to the interval between 'agree' and 'undecided' or equal to the interval between 'disagree' and 'strongly disagree'.

The correlations are presented in Tables 3 and 4. The reader should note that we considered a correlation to be significant when $p < 0.0005$, following the reasoning explained earlier for choice of Type I error rate for this study.

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 0.693 to 0.813. 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.

First Principles of Instruction combined. In Table 4 it can be seen that the combined First Principles scale correlated very highly with ALT ($\rho = 0.682$, $p < 0.0005$, $n = 111$, $\rho^2 = 0.47$), with student Achievement ($\rho = 0.823$, $p < 0.0005$, $n = 110$, $\rho^2 = 0.68$), with Satisfaction ($\rho = 0.830$, $p < 0.0005$, $n = 112$, $\rho^2 = 0.69$), and the overall Instructor/Course Rating (BEST) ($\rho = 0.867$, $p < 0.0005$, $n = 112$, $\rho^2 = 0.75$). The correlation with student Mastery was less strong but nonetheless highly significant ($\rho = 0.341$, $p < 0.0005$, $n = 113$, $\rho^2 = 0.12$). Note that the Class Rating and BEST Rating are indicators of perceived overall quality by students and are correlated very highly ($\rho = 0.799$, $p < 0.0005$, $n = 134$, $\rho^2 = 0.64$).

Academic Learning Time (ALT). It can be seen that ALT is correlated significantly with Learning Achievement ($\rho = 0.602$, $p < 0.0005$, $n = 128$, $\rho^2 = 0.38$). Further indicators of student achievement include the student's Mastery Level. ALT is positively correlated with Mastery Level ($\rho = 0.470$, $p < 0.0005$, $n = 136$, $\rho^2 = 0.22$) and with Course Grade ($\rho = 0.463$, $p < 0.0005$, $n = 126$, $\rho^2 = 0.21$ [not shown in Table 4]).

The ALT scale is also correlated with the BEST Rating ($\rho = 0.605$, $p < 0.0005$, $n = 134$, $\rho^2 = 0.37$), with overall Class Rating ($\rho = 0.496$, $p < 0.0005$, $n = 135$, $\rho^2 = 0.25$), and with Learner Satisfaction ($\rho = 0.515$, $p < 0.0005$, $n = 132$, $\rho^2 = 0.27$).

These results are very strong as a group. ALT is correlated positively and significantly with student's self-reported learning achievement, which is consistent with past studies of ALT. ALT is also correlated positively and significantly with student Mastery of course learning objectives. This can be interpreted to mean that students who agreed that they frequently engaged successfully in problems and doing learning tasks in a course also agreed that they mastered course objectives and received a high grade. Furthermore, they agreed that this was an excellent course and that they were very satisfied with it.

Perhaps even more important for this study, is the evidence supporting the 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. This is strong empirical support for Merrill's claim that student learning is promoted when First Principles of Instruction are used. These relationships will be clarified in the pattern analysis results described below (APT).

It appears that from a student's perspective, when First Principles are used in a course, this is associated with high quality instruction with which they are very satisfied and from which they learned a lot. The reader should note that respondents were unlikely to know that we were measuring First Principles of Instruction, since we never told them or their instructors. We just told them what was on the study information sheet: "The purpose of this study is to examine relationships among items on course evaluation forms, instructional practices, and student academic learning time."

There are many other highly significant and strong correlations in Tables 3 and 4. Space precludes further discussion here. However, this led us to investigation of specific patterns that illuminate further the nature of these relationships.

Pattern Analysis (APT)

The Spearman ρ coefficients indicate correlations of ranks of ordinal measures. While there were numerous highly significant relationships which explained typically between 40 and 70 percent of the variance in ranks, the specific patterns that show temporal relations among 3 or more variables is not shown. 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 (i.e., Achievement)? This is a temporal pattern. Something happened during the course (instructor did stuff, and so did students) and then some kind of outcome later occurred (i.e., students agreed that they learned a lot, or they did not).

Linear models, such as multiple regression analysis, have limitations if one assumes that relations are not linear but instead temporal. Analysis of Patterns in Time (APT) is one way of approaching data analysis that is an alternative to the linear models approach (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).

APT has been used successfully in a past studies. Frick (1990) found, for example, that when "direct instruction" occurred during academic activities in elementary classrooms, the probability of student engagement was 0.97 on average. On the other hand, when "non-direct instruction" occurred, the probability of student engagement was 0.57. The probabilities of these temporal patterns were based on nearly 15,000 one-minute time samples of 25 target students and their school environments. Each target student was observed 8 to 10 hours by trained observers over several months. "Direct instruction" included instructor moves during academic activities that included explaining, demonstrating, questioning, giving feedback, and giving directions. Frick concluded that mildly handicapped students in elementary school settings were 13 times more likely to be off-task during academic activities *if* non-direct instruction was occurring (e.g., student seatwork monitored by the teacher), compared to when direct instruction was occurring.

As another example of research using APT, An (2003) investigated conditions of mode errors when people use modern software. Mode errors occur when the same user action results in more than one outcome, depending on the context. Mode errors can cause serious problems for software users, such as inadvertent destruction of important work, decreased productivity, and task incompletion. Sixteen college students were each asked to perform eight computer tasks during usability tests of three modern direct-manipulation software interfaces. Qualitative analysis of An's results indicated three major types of mode errors: A) right action, wrong result; B) it isn't there where I need it; and C) it isn't there at all. One of the clear temporal patterns that An discovered was: If the mode error was Type A (right action, wrong result), and if the source of the mode error is unaffordance (function not obvious), then the consequence was that users could not find a hidden function or thought they were successful when in fact they were not. An (2003) reported that when the antecedent conditions were true, the consequent event occurred 67 percent of the time. This is a temporal pattern. When users make mode errors of Type A and those errors are due to unaffordance in the computer interface, then about 2 out of 3 times they are unable to find the function they need or they believe mistakenly that they have successfully completed a task.

In our present study, we wanted to know that if ALT and First Principles occur, then what is the likelihood that students will learn a lot, master course objectives, or feel satisfied with their instruction? We did not observe students in college classrooms in our study. We asked them to report on their experiences.

We were able to do APT by using features of SPSS 14 as follows: First, we created new dichotomous variables from existing scale scores for each of the cases.[6] We decided to code a scale as "Yes" if the scale score for that case was greater than or equal to 3.5, and "No" if less than 3.5. Thus, for ALT Agreement if the 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. Other scales were coded similarly for Learning Achievement Agreement, Learner Satisfaction Agreement, First Principles Agreement, and Outstanding Course/Instructor (BEST) Agreement.

If *ALT* **and** *First Principles,* **then** *Learner Achievement*. 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 Learning Achievement 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. Nonetheless, we can estimate the likelihood of a pattern. In SPSS 14, there is a "Tables of Frequencies" procedure that will allow counting of patterns when there is only one observation per variable per case.

Insert Table 5 here

It can be seen in Table 5 that there were a total of 65 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 Learning Achievement is Yes), "followed" in 64 out of those 65 cases, which yields a probability estimate of 64/65 or 0.985.

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 Learning Achievement is Yes? It can be seen that the antecedent occurred a total of 22 times, and the consequent occurred in 7 out of those 22 cases, for a probability estimate of 7/22 = 0.32. Thus, about 1 out of 3 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 more than 3 times as likely (0.98/0.32 = 3.06) to agree that they learned a lot in the course, compared to when ALT and First Principles do not occur.

APT results do not imply causation but temporal association. These kinds of results are similar to epidemiological findings in medicine. For example, 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.

[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.

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 63 out of 66 cases when the antecedent was true for a probability estimate of 0.95. On the other hand, when ALT was No and First Principles was No, then Learner Satisfaction occurred in 6 out of 23 cases, or a probability estimate of 0.26. The estimated odds of Learner Satisfaction when both ALT and First Principles are present compared to when both are not are about 3.6 to 1 (0.95/0.26).

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 62/65 = 0.95. If both antecedent conditions are false, the probability is 5/25 = 0.20. The odds are 0.95/0.20, or about 4.75 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 24 out of 66 students for a probability of 24/66 = 0.36 (roughly 1 out of 3 students). On the other hand, only 1 out of 25 students agreed that they had mastered course objectives (probability = 1/25 = 0.04) when they did not agree that First Principles and ALT occurred. Thus, students were 9 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.

It is interesting to note that 16 out of 25 students considered themselves to be Partial Masters of course objectives when they disagreed that ALT and First Principles occurred. Thus, about 64 percent (16/25) of students report that they can achieve some objectives in the absence of ALT and First Principles. It would seem that college students can learn despite poor instruction, which is not surprising, since students who make it to college tend to have higher aptitude for learning that those who do not (e.g., as evidenced by predictive validity of SAT and GRE scores). The reader should also note that about 62 percent of all 140 respondents indicated they were partial masters (see Table 1).

Factor Analysis: Instructional 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 'Instructional Quality'. Researchers have noted from past studies of course evaluation ratings that they tend to be correlated with each other (e.g., Kulik, 2001), sometimes referring to this as a "halo effect". In other words, if a student is happy with an instructor or course, then he or she tends to rate everything perceived as positive about the course very highly, and vice-versa.

To investigate this possibility, we conducted a factor analysis of the scales reported in Table 4. We used the extraction method called image analysis with varimax rotation. 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). Unique variance of variables is excluded, and only shared or common variance is factor analyzed. The image scores for each case are the predicted scores that are obtained from multiple regressions with other variables in the set, resulting in a variance-covariance matrix with unique variance removed. The principal components method is then used with this matrix instead of the one derived from raw scores. The net effect of this is to minimize the impact of outliers (e.g., see Veldman, Kaufmann, Agard & Semmel (1985), p. 55, for further details).

Results of factor analysis are presented in Table 9. A single factor was extracted with an eigenvalue of 4.88, which accounted for nearly 70 percent of the variance. Remaining factors had eigenvalues less than one. Since only one factor was extracted, no rotation of the solution was done. The factor loadings in Table 9 can be interpreted as correlations of each scale with the overall factor. Loadings ranged from 0.94 for learner Satisfaction to 0.38 for Mastery Level.

Insert Table 9 here

Given that only a single factor could be extracted, this means that all of these scales are strongly associated with a single construct. We shall call this construct 'Instructional Quality'. This means that Instructional Quality was greater when students were satisfied with the instruction, they perceived the course and instructor to be outstanding overall, they agreed that First Principles of Instruction occurred, they were more likely to classify this course as "great", they agreed that they frequently engaged *successfully* in problems, assignments and tasks related to course objectives (ALT), and that they were more likely to report that they had mastered those objectives.

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. APT measures the relation. APT probability estimates are relatively easy to comprehend and can have practical implications.

For example, one can choose to minimize the chances of getting lung cancer by refraining from smoking cigarettes and avoiding smoky environments. The U.S. Surgeon General started mandating warnings on cigarette packs in the 1960s, long before causation was established from controlled studies. The pattern was clear to physicians back then, even though it was a temporal relationship.

Instructors could choose to facilitate student Academic Learning Time and incorporate First Principles of Instruction in their courses. If they do, then results from our study predict that students will be more satisfied, achieve more, and rate such courses more highly.

Instructors can test this prediction by using the scales from this study to evaluate a course currently taught. Consider this evaluation as a baseline measure. Then instructors can try modifying their courses and actions to increase ALT and use First Principles throughout the course. Then use the scales from this study to evaluate the new version of the course. Do the ratings improve for this redesigned course? At the same time, consider another course taught as a baseline, evaluate it, do not change it, and then evaluate it the next time it is taught. Do the ratings stay about the same for this unmodified course?

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.

An astute reader will notice that APT results could have been accomplished with cross tabulation. This is true when there is one observation per variable per case, variables are coded at a nominal level, and there is a theoretical or practical reason for assuming temporal order even though it is not specifically represented in the data. However, the possibilities for query formation for temporal patterns in APT go well beyond what simple crosstabulation can do in SPSS. See Frick (1983) for examples of query syntax and counting rules. APT&C software is currently under development by Frick (2005).

Mastery of learning objectives. As noted earlier, only 1 out of 4 students considered themselves masters of course objectives, even though 87 percent received A's and B's for their course grades.

Furthermore, 79 percent agreed or strongly agreed that they learned a lot during a course compared to when they began it. And while there is a significant statistical association between Learning Achievement and Mastery of course objectives, it is clear that students could be learning more in their courses.

This brings us back to 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.

While in our study we asked students about their mastery of course objectives, not about their literacy proficiency, the trend is similar – and not a good trend. Those critics of our U.S. education system who claim that students coming out of our postsecondary schools lack necessary skills are supported by results from our study. Data from this study are consistent with findings from Baer, Cook and Baldi (2006). Even though we were not able to randomly sample students in our study, those students rated a wide range of courses and topics (at least 89 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. Findings from this study are the strongest thus far which provide empirical support for First Principles of Instruction (D. Merrill, personal communication). 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 Satisfaction and Learning Achievement (Kirkpatrick's Levels 1 and 2), Academic Learning Time (ALT) (Berliner, 1991; Fisher, et al., 1978), and global course and instructor ratings (cf., Cohen, 1981; Kulik, 2001). Data from our study indicate that these rating scales are highly reliable, ranging from 0.74 to 0.97. The validity of these scales is supported by the high intercorrelations among them. These are strong correlations and highly significant statistically.

Thus, these scales should be given serious consideration for implementation in course evaluations of college instructors. Unlike many items on typical course evaluations used, these scales have a demonstrated relationship with outcome measures such as self-reports of student satisfaction and learning achievement. 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. As discussed above, instructors can carry out their own experiments. If they increase use of First Principles in classes, do their ratings also increase? Do more students achieve mastery of course objectives? Mastery can be assessed by objective measures, not solely by asking students to self-report on their accomplishments.

Most importantly, when instructors consider use of First Principles of Instruction, this requires identification of authentic problems for students to solve – real-life problems. Such problems should be more motivating and interesting for students, and if they are more motivated, then they may 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). Meaning of learning is further enhanced when students can integrate what they have learned into their personal lives. In other words, what they are learning is relevant – not just some requirement by their instructors, or hoops to jump through to complete a course. 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 instructor were outstanding ("really great", $\rho = 0.87$).

Conclusion

We surveyed 140 undergraduate and graduate students from at least 89 different courses at several universities. Scale reliabilities ranged from 0.74 to 0.97. Correlations among scales averaged 0.63 and were highly significant at $p < 0.0005$. The overall Type I error rate for our study, in which 58 statistical tests were performed, was $p < 0.0286$. Thus, chances are about 3 in 100 that a statistical test would produce a significant result when no relationship exists. Of the 58 statistical tests we performed, 21 of them were not significant. There was no statistically significant association between gender and each of the 9 scales, nor among 12 comparisons of demographic variables themselves ($p > 0.0005$).

Results from analysis of patterns in time indicated that students were 3-5 times more likely to learn a lot and were satisfied with courses when first Principles of Instruction were used *and* students were frequently engaged successfully (ALT). Students were 9 times more likely to master course objectives when both First Principles and ALT occurred, compared with their absence.

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 try. 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.

Limitations and Further Research

This was a correlational study. Students were volunteers and provided self-reports on what happened in their courses. Correlation does not imply causation, nor do we know for sure why some students elected to complete the survey and others did not.

Since there was no way that their grades received could be affected by their evaluation in this study, the findings are not compromised in this way. Data were collected via reliable methods and stored in a location on a Web server account only accessible by the first author.

What we do know is that we asked instructors at several institutions to ask their students to complete the survey instrument, and we also asked students through campus organizations. We do know that we have a very wide representation of course topics and instructors. We have no reason to believe that students were fabricating responses. The qualitative data made sense to the investigators. Responses were unique from one case to the next, and the language used to describe their courses and instructors appeared to be typical in our experiences of dealing with students at the college level. There did not appear to be any “ringers” who were trying to influence the outcome of the study. In fact, there was no evidence that the same individual completed the survey more than once.

Moreover, many studies in the past have shown good correlations with self-reports and student ratings when compared with other measures. Thus, we do not believe that because respondents were volunteers that the findings are biased in any particular way. Arguing against the possibility of bias are the wide range of courses and instructors evaluated, students at all levels of their education, and the consistency of our findings with those from other studies.

The one statistic that is noteworthy is that nearly 58 percent of the classes were rated as ‘great’. We expected the modal category to be ‘about average’. This rating of ‘great’ is consistent with the percent of respondents who agreed or strongly agreed with many of the items (as seen in Table 2), in which they were about twice as likely to agree as not. If anything, then, the majority of the courses evaluated in this study were considered excellent by students. Ratings of courses that were ‘about average’ or ‘really awful’ may be underrepresented in our sample. Nonetheless, there was sufficient variation in ratings of excellent and poor courses that we found strong correlations and obtained highly reliable measurement scales from this sample. Had there not been such variation, then scale reliabilities would have been poor and correlations would have been attenuated by low reliabilities. Clearly that was not the case in this study.

The next kind of research that needs to be done is to obtain independent external measures of many of the *TALQ Scales*. For example, classroom observations could be made on student Academic Learning Time in courses. Similarly, classroom observations could be made on use of First Principles of Instruction. Then these observations could be compared to student ratings of the same factors.

Achievement could be measured by pre- and posttests to see how these learning gains compare with student reports on their own evaluations of whether they learned a lot. When these kinds of validation studies have been done in the past, as reported in numerous meta-analyses (cf., Cohen, 1981; Feldman, 1989; Kulik, 2001), student ratings have been found to correlate well with external measures. We would expect similar results for the *TALQ Scales*, but this needs to be investigated by further research.

Finally, as alluded to above, instructors can use the *TALQ Scales* to conduct their own classroom experiments. When instructors increase use of First Principles and students increase their ALT, do Satisfaction, Achievement and perceived Instructor/Course Quality also increase? When instructors do not increase those factors, do those outcomes not increase? If these patterns obtain, then this is further evidence that First Principles and Academic Learning Time make a real difference in quality of instruction and student achievement.

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.

An, J. (2003). *Understanding mode errors in modern human-computer interfaces: Toward the design of usable software*. Bloomington, IN: Ph.D. dissertation.

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.

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 an extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30*, 583-645.

Ferguson, G. (1971). Statistical analysis in psychology and education (3rd ed.). New York, NY: McGraw-Hill.

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. (2005). Bridging qualitative and quantitative methods in educational research: Analysis of patterns in time and configuration (APT&C). Proffitt Grant Proposal. Retrieved March 4, 2007: http://education.indiana.edu/~frick/proposals/apt&c.pdf .

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 .

Frick, T. & Semmel, M. (1978). Observer Agreement and Reliabilities of Classroom Observational Measures. *Review of Educational Research, 48*(1), 157-184.

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.

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.

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. Descriptive statistics on respondent and course demographics

		Frequency	Percentage
Gender	Female	93	66.4
	Male	43	30.7
	Missing	4	2.9
	Total	140	100.0
Class Rating: I would rate this class as:	Great	81	57.9
	Average	44	31.4
	Awful	13	9.3
	Missing	2	1.4
	Total	140	100.0
Expected Grade: In this course, I expect to receive (or did receive) a grade of:	A	92	65.7
	B	30	21.4
	C	6	4.3
	D	1	0.7
	N/A or Don't know	10	7.1
	Missing	1	0.7
	Total	140	99.9
Achievement: With respect to achievement of objectives of this course, I consider myself a:	Master	35	25.0
	Partial master	87	62.1
	Nonmaster	17	12.1
	Unknown	1	0.7
	Total	140	99.9
Class Standing: I am a:	Freshman	23	16.4
	Sophomore	19	13.6
	Junior	23	16.4
	Senior	19	13.6
	Graduate	48	34.3
	Other	7	5.0
	Missing	1	0.7
	Total	140	100.0
Course Setting: I took this course:	Face to face	97	69.3
	Blended	8	5.7
	Online	34	24.3
	Missing	1	0.7
	Total	140	100.0

Table 2.1. Percentages of Respondents: Academic Learning Time Scale (α = 0.85)

Item No.	*Item Stem*	*SD*	*D*	*U*	*A*	*SA*	*Valid N*
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.~~	58.6	22.1	6.4	5.0	2.1	132
12	I frequently did very good work on projects, assignments, problems and/or learning activities for this course.	1.4	3.6	10.0	37.9	40.0	130
14	I spent a lot of time doing tasks, projects and/or assignments, and my instructor judged my work as high quality.	2.9	10.7	16.4	40.7	20.0	127
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.	3.6	9.3	16.4	37.9	29.3	135
29-	~~I did a minimum amount of work and made little effort in this course.~~	56.4	26.4	9.3	3.6	2.1	137

Table 2.2. Percentages of Respondents: Learning Achievement Scale (α = 0.97)

Item No.	*Item Stem*	*SD*	*D*	*U*	*A*	*SA*	*Valid N*
4	Compared to what I knew before I took this course, I learned a lot.	2.1	5.7	8.6	32.9	48.6	137
10	I learned a lot in this course.	4.3	3.6	12.1	30.7	48.6	139
22	Looking back to when this course began, I have made a big improvement in my skills and knowledge in this subject.	3.6	8.6	14.3	30.7	40.0	136
27-	I learned very little in this course.	50.7	26.4	8.6	8.6	3.6	137
32-	I did not learn much as a result of taking this course.	49.3	25.7	10.0	8.6	4.3	137

Table 2.3. Percentages of Respondents: Items selected from BEST standard university form (α = 0.92)

Item No.	*Item Stem*	*SD*	*D*	*U*	*A*	*SA*	*Valid N*
8	Overall, I would rate the quality of this course as outstanding.	10.7	9.3	9.3	29.3	40.0	138
13	~~This course is one of the most difficult I have taken.~~	12.9	42.9	16.4	19.3	7.9	139
16	Overall, I would rate this instructor as outstanding.	8.6	8.6	11.4	22.9	44.3	134
18	~~This course increased my interest in the subject matter.~~	7.1	10.0	11.4	35.0	36.4	140
38	Overall, I would recommend this instructor to others.	9.3	7.1	11.4	19.3	47.9	133

Table 2.4. Percentages of Respondents: Authentic Problems Scale (α = 0.81)

Item No.	*Item Stem*	*SD*	*D*	*U*	*A*	*SA*	*Valid N*
3	I performed a series of increasingly complex authentic tasks in this course.	6.4	7.9	20.0	36.4	26.4	136
19	~~My instructor directly compared problems or tasks that we did, so that I could see how they were similar or different.~~	5.0	14.3	14.3	27.1	31.4	129
25	I solved authentic problems or completed authentic tasks in this course.	2.9	8.6	10.7	42.9	31.4	135
31	In this course I solved a variety of authentic problems that were organized from simple to complex.	2.1	15.7	17.1	33.6	25.0	131
33	Assignments, tasks, or problems I did in this course are clearly relevant to my professional goals or field of work.	2.1	10.0	15.1	26.4	44.3	138

Table 2.5. Percentages of Respondents: Activation Scale (α = 0.91)

Item No.	*Item Stem*	*SD*	*D*	*U*	*A*	*SA*	*Valid N*
9	I engaged in experiences that subsequently helped me learn ideas or skills that were new and unfamiliar to me.	4.3	6.4	7.9	38.6	41.4	138
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.	3.6	7.9	13.6	42.1	28.6	134
30	My instructor provided a learning structure that helped me to mentally organize new knowledge and skills.	8.6	12.1	11.4	30.7	34.3	136
39	In this course I was able to connect my past experience to new ideas and skills I was learning.	4.3	11.4	11.4	32.9	34.3	132
41-	In this course I was not able to draw upon my past experience nor relate it to new things I was learning.	43.6	22.9	14.3	12.1	2.1	133

Table 2.6. Percentages of Respondents: Demonstration Scale (α = 0.88)

Item No.	*Item Stem*	*SD*	*D*	*U*	*A*	*SA*	*Valid N*
5	My instructor demonstrated skills I was expected to learn in this course.	5.0	5.0	13.6	28.6	42.1	132
15	~~Media used in this course (texts, illustrations, graphics, audio, video, computers) helped me to learn instead of distracting me.~~	5.0	9.3	12.9	36.4	30.7	132
17	My instructor gave examples and counter-examples of concepts that I was expected to learn.	2.9	10.0	10.0	35.7	35.7	132
35-	My instructor did not demonstrate skills I was expected to learn.	44.3	25.0	10.0	10.0	3.6	130
43	My instructor provided alternative ways of understanding the same ideas or skills.	6.4	10.7	15.7	30.0	29.3	129

Table 2.7. Percentages of Respondents: Application Scale (α = 0.74)

Item No.	*Item Stem*	*SD*	*D*	*U*	*A*	*SA*	*Valid N*
7	My instructor detected and corrected errors I was making when solving problems, doing learning tasks or completing assignments.	5.7	12.9	10.0	29.3	30.7	124
23	My instructor gradually reduced coaching or feedback as my learning or performance improved during this course.	7.1	23.6	29.3	18.6	9.3	123
26-	~~Opportunities to practice what I learned during this course (e.g., assignments, class activities, solving problems) were not consistent with how I was formally evaluated for my grade.~~	26.4	29.3	20.7	11.4	7.1	133
36	I had opportunities to practice or try out what I learned in this course.	2.1	10.0	13.6	40.0	30.0	134
42	My course instructor gave me personal feedback or appropriate coaching on what I was trying to learn.	7.1	10.0	12.9	29.3	35.0	132

Table 2.8. Percentages of Respondents: Integration Scale (α = 0.81)

Item No.	*Item Stem*	*SD*	*D*	*U*	*A*	*SA*	*Valid N*
11	I had opportunities in this course to explore how I could personally use what I have learned.	3.6	7.9	12.1	35.0	39.3	137
28	I see how I can apply what I learned in this course to real life situations.	2.9	5.7	11.4	36.4	43.6	140
34	I was able to publicly demonstrate to others what I learned in this course.	3.6	12.1	17.9	33.6	27.1	132
37	In this course I was able to reflect on, discuss with others, and defend what I learned.	4.3	10.0	15.7	35.7	26.4	129
44-	I do not expect to apply what I learned in this course to my chosen profession or field of work.	52.1	24.3	9.3	10.7	2.1	138

Table 2.9. Percentages of Respondents: Learner Satisfaction Scale (α = 0.94)

Item No.	*Item Stem*	*SD*	*D*	*U*	*A*	*SA*	*Valid N*
2	~~I am very satisfied with how my instructor taught this class.~~	10.7	10.0	10.7	22.9	42.1	135
6-	I am dissatisfied with this course.	54.3	21.4	7.1	7.1	7.9	137
20-	This course was a waste of time and money.	54.3	25.0	7.9	5.7	6.4	139
40	~~I enjoyed learning about this subject matter.~~	2.9	6.4	12.9	34.3	41.4	137
45	I am very satisfied with this course.	8.6	7.9	12.1	22.9	45.7	136

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

		Authentic Problems	Activation	Demonstration	Application	Integration
Authentic Problems	ρ	1.000				
	N	137				
Activation	ρ	.738**	1.000			
	N	127	128			
Demonstration	ρ	.735**	.769**	1.000		
	N	123	118	124		
Application	ρ	.760**	.693**	.740**	1.000	
	N	136	127	123	138	
Integration	ρ	.812**	.813**	.737**	.714**	1.000
	N	133	125	122	134	135

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

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

		First Principles	ALT	Achievement	Satisfaction	Mastery	Class Rating	BEST Rating
First Principles	ρ	1.000						
	N	114						
ALT	ρ	.682**	1.000					
	N	111	137					
Achievement	ρ	.823**	.602**	1.000				
	N	110	128	131				
Satisfaction	ρ	.830**	.515**	.874**	1.000			
	N	112	132	128	135			
Mastery	ρ	.341**	.470**	.301**	.361**	1.000		
	N	113	136	130	134	139		
Class Rating	ρ	.735**	.496**	.760**	.853**	.319**	1.000	
	N	112	135	129	133	138	138	
BEST Rating	ρ	.867**	.605**	.759**	.859**	.386**	.799**	1.000
	N	112	134	128	132	135	134	136

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

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

	ALT Agreement			
	No		Yes	
	First Principles Agreement		First Principles Agreement	
	No	Yes	No	Yes
	Learner Achievement Agreement	Learner Achievement Agreement	Learner Achievement Agreement	Learner Achievement Agreement
	Count	Count	Count	Count
No	15	1	5	1
Yes	7	9	5	64

Table 6. APT Results for the Pattern: If *ALT* and *First Principles,* then *Learner Satisfaction*

	ALT Agreement			
	No		Yes	
	First Principles Agreement		First Principles Agreement	
	No	Yes	No	Yes
	Learner Satisfaction Agreement	Learner Satisfaction Agreement	Learner Satisfaction Agreement	Learner Satisfaction Agreement
	Count	Count	Count	Count
No	17	2	7	3
Yes	6	8	3	63

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

	ALT Agreement			
	No		Yes	
	First Principles Agreement		First Principles Agreement	
	No	Yes	No	Yes
	BEST Agreement	BEST Agreement	BEST Agreement	BEST Agreement
	Count	Count	Count	Count
No	20	2	9	3
Yes	5	8	1	62

Table 8. APT Results for the Pattern: If *ALT* and *First Principles,* then *Mastery*

	ALT Agreement			
	No		Yes	
	First Principles Agreement		First Principles Agreement	
	No	Yes	No	Yes
	With respect to achievement of objectives of this course, I consider myself a:	With respect to achievement of objectives of this course, I consider myself a:	With respect to achievement of objectives of this course, I consider myself a:	With respect to achievement of objectives of this course, I consider myself a:
	Count	Count	Count	Count
Nonmaster	8		1	3
Partial Master	16	9	6	39
Master	1		3	24

Table 9. Factor Matrix for Main Scales

	Factor
	1
Satisfaction	.945
BEST Rating	.920
First Principles	.906
Achievement	.897
Class Rating	.891
ALT	.505
Mastery Level	.381