Dependability of College Student Ratings of Teaching and Learning Quality

Rajat Chadha, Ph.D.

Australian Council for Educational Research

Theodore W. Frick, Ph.D.

School of Education, Indiana University Bloomington

Paper presented at the annual conference of the
American Educational Research Association,

New Orleans, April 9, 2011

Abstract

The U.S. Commission on the Future of Higher Education (2006) has expressed concerns about the quality of higher education. As a response to accountability demands, such as those of the Commission, several institutional-level assessment efforts have been undertaken, such as administration of standardized tests and surveys of student engagement. Nonetheless, universities and colleges have often used student ratings as a means of evaluating courses and instruction—because they are very practical. According to meta-analyses, only a few items on typical course evaluations have been found to be related to student achievement. To address these concerns, Frick, Chadha, Watson, Wang and Green (2009) developed a new course evaluation instrument that consists of nine student rating scales of teaching and learning quality (TALQ). Five TALQ scales measure *First Principles of Instruction*, which were derived from a synthesis of instructional theories by Merrill (2002; 2009). Other TALQ scales measure student perceptions of successful engagement (academic learning time) and learning progress. The present study examined the dependability of TALQ scale scores, since these psychometric properties have not been previously addressed.

The TALQ was administered near the end of the semester to 464 students in 12 classes taught by 8 professors at a large Midwestern university. Results of generalizability studies revealed that TALQ scale scores were dependable for student ratings of overall course and instructor quality, satisfaction, learning progress and on 3 *First Principles of Instruction* (activation, application and integration). Two of the *First Principles* scale scores were found to be less dependable (authentic problems and instructor demonstration), as well as student academic learning time. Based on these findings, changes in 3 items are suggested. Changes in 5 items on the academic learning time scale are also recommended. Future validation studies are

recommended that investigate the use of the modified TALQ scales for improving the quality of teaching in postsecondary education.

1. Introduction

The U.S. Commission on the Future of Higher Education (2006) has expressed concerns about the quality of higher education, especially student learning. As a response to accountability demands, such as those of the Commission, several institutional-level assessment efforts have been undertaken, such as administration of standardized tests and surveys of student engagement. Feedback on how to improve student learning is best available at the course level rather than at the institution level, especially considering the diversity of courses and skills required for performance in those courses. Course evaluations have the potential to play a major role, if feedback from them could be used to inform what aspects of *teaching* need to be improved—aspects which are associated with improved student learning outcomes.

Universities and colleges have often used student ratings as a means of evaluating courses and instruction. However, according to meta-analyses (Cohen, 1981; Feldman, 1989), only a few items on typical course evaluations have been found to be related to student achievement. Some of these items are the global instructor and course quality items. Scores on these items have positive association with student learning achievement. However, feedback from these scores is apparently not useful in improving teaching. Researchers (L'Hommedieu, Menges & Brinko, 1990; Lang & Kersting, 2007; Spencer & Flyr, 1992) on faculty use of course evaluations for improving instruction have generally found:

1. Only short-term gains (with small-to-modest effect size) at best in overall student ratings resulted from using course evaluation feedback.

2. No long-term gains in overall student ratings resulted from using course evaluation feedback.
3. Course evaluation feedback was not used to improve teaching.

Therefore, there is an apparent need for a better course evaluation instrument, which not only provides useful feedback that can help in improving teaching, but also yields reliable and valid scores to be used as an indicator of teaching effectiveness. Toward this end, Frick, et.al. (2009) developed a new course evaluation instrument, consisting of Teaching and Learning Quality (TALQ) scales. If the TALQ were to successfully predict student learning achievement, then instructors who receive low ratings on TALQ scales might be more motivated to modify their courses and teaching/learning activities—particularly if universities were to adopt TALQ scales and use them for assessing quality of teaching in merit review, tenure and promotion.

1.1 Teaching and Learning Quality (TALQ) Course Evaluation

The Teaching and Learning Quality scales for course evaluation attempt via student ratings to measure student academic learning time (Berliner, 1991; Rangel & Berliner, 2007), instructor use of *First Principles of Instruction* (Merrill, 2002; 2007; 2009; Merrill, Barclay, & van Schaak, 2008), overall instructor and course quality, satisfaction with the course (Kirkpatrick, 1994), and students' perceptions of personal learning progress. Frick et al. (2009) created the TALQ scales based on extant theory and empirical evidence on how these factors help promote student learning. These measures are described below in more detail.

1.1.1 Academic Learning Time

Student Engagement. Student engagement has been defined as students' time and energy invested in the pursuit of learning in and out of the class (Newmann, Wehlage, & Lamborn, 1992; Finn 1989; Fredricks, Blumenfeld, & Paris, 2004). Student engagement has been a much-

studied subject of research in the literature. In a longitudinal study of students enrolled in several institutions granting bachelor‘s degrees, Astin (1993) reported positive relationship between student engagement in academic activities and learning achievement. In another report, Kuh, Kinzie, Buckley, Bridges and Hayek (2006) found that student engagement is positively associated with student learning achievement and students' overall academic development in higher education. Considering the findings from these studies and the potential of teachers to change student engagement (Finn & Rock, 1997; Kuh, 2001, Kuh, 2003), Frick et.al. (2009) considered it worthwhile to study student engagement.

Academic Learning Time. ALT refers to frequency of successful student engagement in learning activities relevant to curriculum goals (Berliner 1991; Brown & Saks, 1986; Fisher et al., 1978; Kuh, et al., 2006; Squires, Huitt & Segars, 1983). Past research has demonstrated that ALT is positively correlated with student learning achievement (Berliner, 1991; Rangel & Berliner, 2007; Fishe et al., 1978). Moreover, it was found that increased ALT was negatively correlated with negative attitude toward school, mathematics, and reading (Fisher et al.,1978).

Student Engagement vs. Academic Learning Time. Engaged time only accounts for the time students are on-task. ALT takes this a step further and accounts for the time and frequency with which students are engaged in relevant tasks *successfully*. Therefore, Frick et al. (2009) attempted to measure academic learning time on the TALQ course evaluation rather than engaged time.

ALT depends in part on student effort and is not under direct control of a teacher. ALT is a better predictor of student learning achievement in the research literature than is student engagement time. Although instructors should not be held accountable for student ALT, it is nonetheless a good predictor. For example, Frick, Chadha, Watson and Zlatkovska (2010) found

that students who agreed that they experienced ALT were more likely to be rated by their instructors as having achieved a high level of student mastery by a factor of 3 to 1. Students who agreed that they experienced ALT were three times more likely to be high masters of course objectives, compared with students who did not agree, according to instructor evaluations independently obtained after the course was over.

Thus, successful student engagement that occurs frequently (ALT) is a predictor of student learning achievement, even though instructors have no direct control over student ALT. Are there things over which instructors do have control, which in turn are likely to predict increased ALT?

1.1.2. First Principles of Instruction

Merrill (2002; 2007; 2009) synthesized five prescriptive instructional design factors after an extensive review of instructional design theories, models, and empirical research (e.g., Tennyson, Schott, Seel, & Dijkstra, 1997; Gagne, 1985; Glaser, 1992; Marzano et al., 2001; McCarthy, 1996; Reigeluth, 1983, 1987, 1999; Tennyson et al., 1997; van Merriënboer, 1997) Merrill found that one or more of the factors were present in each of these models and theories. He called these factors *First Principles of Instruction*. These First Principles include: 1) authentic problems (students solve a series of increasingly complex real-world problems, or complete authentic whole tasks); 2) activation (students link their past learning or experience to 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). Merrill claimed that: 1) to the extent that these principles are present during instruction, learning is promoted; 2) these

principles are general and can be implemented in any type of instructional delivery system (online or face to face instruction) or instructional architecture (direct instruction, tutorials, experiential methods, exploratory methods, etc.); and 3) these principles are prescriptive (how instruction is designed to promote learning activities) rather than learning-oriented (how students learn) (Merrill, 2002; 2007; 2009).

Research on First Principles of Instruction. Meta-analyses of empirical research on effective teaching that have identified instructional techniques that support one or more of the *First Principles of Instruction* include: Ellis and Worthington (1994) (activation, demonstration, application, and authentic problems principles); Friedman and Fisher (1998) (activation, application, and integration principles); Marzano, Pickering and Pollock (2001) (authentic problems, activation, demonstration, application, and integration principles); and Marzano (2003) (activation, demonstration, and application principles).

Empirical studies of all five First Principles of Instruction. One experimental study has investigated the relationship between the all five *First Principles of Instruction* approach (taken as a whole) and student learning achievement. A study was conducted at Thompsom/NETg (Thomson, 2002), where instructional designers applied the *First Principles of Instruction* to revise an existing e-learning Excel course to a new scenario-based course. On the posttest comprised of three authentic Excel tasks, students enrolled in the course which was designed using the *First Principles of Instruction* evidenced statistically significantly higher learning achievement (mean score: 89%), compared with students enrolled in the existing course (68%) and the control group (34%) who did not receive any instruction.

In a further study of 464 students enrolled in 12 different courses at a large Midwestern university, Frick, Chadha, Watson and Zlatkovska (2010) reported strong to very high Spearman

correlation coefficients ($p < 0.0005$) between student ratings of instructor use of *First Principles of Instruction* and student self-reported academic learning time ($r = 0.583$), student self-reported learning progress ($r = 0.725$), student satisfaction with the course ($r = 0.778$), and student ratings of overall instructor and course quality ($r = 0.774$). They also found that if students agreed that their instructors used *First Principles of Instruction* and they also agreed that they experienced academic learning time (ALT), those students were about 5 times more likely to be rated by their instructors as having achieved a high level of mastery of course objectives—when compared with students who did not agree that their instructors used *First Principles* and who did not agree that they experienced ALT. Perhaps even more significant was the finding that when students did *not* agree that their instructors used *First Principles* and did *not* agree that they experienced ALT, they were about 26 times more likely to be rated at a low level of mastery by their instructors—when compared with students who did agree. Instructors independently evaluated student mastery levels on a 10-point scale after the course was complete and had no knowledge of student course ratings on the TALQ scales. Instructors based their ratings on student performance in the course (e.g., tests, projects, papers, quality of class participation, etc.)

Other Theories and Models in support of the First Principles of Instruction. Gagné (Gagné, 1965; Gagné& Briggs, 1974; Gagné, Briggs, & Wager, 1988; Gagné, Wager, Golas, & Keller, 2004) prescribed nine events of instruction based on the information processing mental model. These nine events are: gain attention, inform learners of objectives, stimulate recall of prior learning, present the content, provide learning guidance, elicit performance (practice), provide feedback, assess performance, enhance retention, and transfer to the job. The 'stimulate recall of prior learning' event is consistent with the *activation principle* from Merrill's *First Principles of Instruction*. The 'present the content' event is consistent with the *demonstration*

principle. Four of Gagné's nine events ('provide learning guidance', 'elicit performance (practice)', 'provide feedback', and 'assess performance') are consistent with the *application principle*. The 'enhance retention and transfer to the job' events are consistent with Merrill's *integration principle*.

Other models/theories that parallel some of the *First Principles of Instruction* include:

- Star Legacy learning cycle for effective instruction, at the Vanderbilt Learning Technology Center (Schwartz, Lin, Brophy, & Bransford, 1999) parallels *activation, demonstration, application,* and *integration principles*;
- Cognitive Training Model (Foshay, Silber and Stelnicki, 2003) parallels *activation, demonstration, application,* and *integration principles*;
- McCarthy's 4MAT model (McCarthy, 1996) parallels *activation, demonstration, application,* and *integration principles*;
- Jonassen's constructivist learning environment (Jonassen, 1999) parallels *authentic problems, activation, demonstration, application,* and *integration principles*;
- 4C/ID instructional design model (van Merriënboer, 1997; van Merriënboer & Kirschner, 2007) parallels *authentic problems, activation, demonstration, application,* and *integration principles*;
- Learning by Doing model (Schank, Berman, & Macperson,1999) parallels *activation, demonstration, application,* and *integration principles*; and
- Sugrue's (2004) instructional design principles parallel *authentic problems, activation, demonstration,* and *application*.

In summary, a number of instructional theories and models appear to support the applicability of the *First Principles of Instruction* in diverse subject areas. Therefore, Frick et al.

(2009; 2010) considered it worthwhile to have students rate their instructor's use of the *First Principles of Instruction* as an indicator of teaching effectiveness. Scales were comprised of indicators of each of these *First Principles*. Students did not know which indicators belonged to which scale.

1.1.3 Student Satisfaction

Kirkpatrick's (1994) four levels of evaluation of training effectiveness have been used for over five decades in non-formal educational settings such as business and industrial training. These four levels of evaluation are: 1) *satisfaction* with the training that is often referred to as a smiles test, 2) *learning achievement*, 3) *transfer* of learning to a trainee's job or workplace, and 4) *impact* on the overall organization to which the trainees belong. Although Kirkpatrick (1994) suggested these levels of evaluation for training programs, levels 1 and 2 are also employed in higher education settings. The end-of-term course evaluations are an example of level 1 evaluation. Level 2 evaluations occur when professors conduct course assessment tests, course examinations, and other assignments in order to assign a grade for the student learning achievement. Levels 3 and 4 are not usually employed in the context of higher education. Items that measure student satisfaction are present on typical course evaluations used in higher education settings. Although these items do not inform instructors what aspects of teaching need to be improved, Frick et al. (2009) decided to include levels 1 and 2 in order to study their relationship with student global rating items and other TALQ scales.

Using the TALQ scales to measure ALT and *First Principles* could be potentially beneficial in assessing and improving the quality of instruction. Feedback on items from these measures has the potential to inform an instructor what aspects of both the course and his or her teaching need to be improved.

2. Purpose of the Study

The purpose of this study was to investigate the measurement properties of the TALQ course evaluation scales. Specifically, we investigated the dependability of student ratings on TALQ scales for a semester-long class. We wanted to identify potential problems in the TALQ course evaluation scales before undertaking advanced validation studies.

2.1 Research Questions

1. Are mean student ratings for a semester-long *class* dependable, as measured by TALQ course evaluation scales established *a priori*?
2. Are mean ratings for *a student* dependable, based on self-reported academic learning time and learning progress scales?

3. Method

Instrument: The first page of the course evaluation included items on gender, overall class rating (I would rate this class as), expected grade, student status (freshman, sophomore, junior, senior, graduate student, other), and mastery of course objectives rating. The next three pages include 40 items belonging to the following nine different *a priori* TALQ scales:

- Academic Learning Time scale
- Learning Progress scale
- Student Satisfaction scale
- Global course and instructor quality items

Use of five *First Principles of Instruction* in the course:

- Authentic Problems scale
- Activation scale
- Demonstration scale

- Application scale
- Integration scale

Items on these nine scales were scrambled into a random order, rather than grouping them by these scales. No information about these scales was provided. Thus, the respondents did not know about the scales and the specific items that belong to each of the scales. Six faculty members at a large Midwestern university reviewed the TALQ instrument after Frick et al. (2009) prepared the first draft. Based on faculty members' feedback, confusing or ambiguous items were modified. Some items on TALQ instrument were negatively worded to see if students were reading the items carefully.

The course evaluation instrument is presented in Appendix A. Items belonging to each of the nine scales are reported in Appendix B. The number next to an item indicates the sequence of that item on the 40-item TALQ course evaluation. Likert scale ratings were reversed for negatively worded items prior to data analysis. Such antithetical items were included in order to detect possible response bias (e.g., circling 'agree' for all items) and to help insure that students were reading the items carefully.

Data Collection: In this study, instructors (faculty members) from various departments were recruited at a large Midwestern research university. E-mail was sent to faculty mailing lists by the director of a teaching center at the university who agreed to help the researchers. Instructors who expressed interest in participating were contacted with details regarding the study. Eight instructors volunteered to participate in the study. The TALQ course evaluation instrument was administered to students in 12 classes taught by these instructors. One of the researchers accompanied the instructor to the class during 13th, 14th, and 15th week of the fall semester to seek student participants. Instructors left the classroom while students completed the

course evaluation. This is the general practice in administering course evaluations at this university.

Each course evaluation form had a unique code number on the cover page and on the first page of the actual evaluation form. Each participating student wrote his or her name on the cover sheet, which was then detached and given to the instructor before he or she left the classroom. This ensured that the researcher did not know the names of the students who participated. The researcher collected the completed course evaluation forms from the students, which contained only code numbers. As a result of using this coding scheme, instructors did not see the individual student ratings and researchers did not see the student names.

After the end of the semester, cover pages were returned to the researchers by the instructors with ratings of student mastery of course objectives. Instructors removed the top halves that contained the student names before they were sent to the researcher. Thus, the researchers could match the code numbers from the instructor ratings with the code numbers on the student rating forms. No class credit was given to students who participated in the study.

Once collected, one of the researchers entered the data in SPSS using a separate file for each class. Another person cross-checked the SPSS data.

Instructor Participants: Eight instructors participated in the study and they taught 12 different classes. Two instructors taught 2 different courses each. One instructor taught 2 sections of one course and 1 section of another course. Five instructors taught 1 course each. These 12 courses were from diverse subject areas: business; philosophy; history; kinesiology; social work; computer science; nursing; and health, physical education, and recreation.

Student Participants: Four hundred and sixty-four students completed the Teaching and Learning Quality (TALQ) course evaluation instrument. Fifty-six percent of the respondents

were female and 44% were male. This is very similar to the gender proportion on this university campus. Almost all of the respondents were undergraduate students (52 freshmen, 104 sophomores, 115 juniors and 185 seniors). A large number of student respondents were juniors or seniors. This student distribution was expected since the university faculty members usually teach advanced courses and associate instructors often teach the introductory courses. Amongst the 12 classes, only one class was at the freshman level (class 2). The number of student respondents who completed the TALQ ranged from 16 to 104 in the 12 classes, though in 10 of the 12 classes the range was from 22 to 53. The response rates in classes ranged from 49% to 100%.

4. Generalizability Theory Study Designs

When investigating the reliability of an instrument, the classical test theory approaches to reliability provide a measure of proportion of the true score variance to the total variance. The two variance components that are estimated when using one of the classical test theory reliability approaches (test-retest, parallel forms, internal consistency) are: true score variance and error variance. Classical test theory reliability approaches can estimate error variance from only one source of error in a single analysis. This shortcoming of classical test theory reliability approaches is addressed by generalizability theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972). Generalizability theory designs permit estimating error variance from more than one source in a single analysis. Generalizability theory informs about the dependability of measurements. Shavelson and Webb (1991) describe dependability as:

> Dependability, then, refers to the accuracy of generalizing from a person's observed score on a test or other measure (e.g., behavior observation, opinion

survey) to the average score that person would have received under all possible conditions that the test user would be equally willing to accept. (p.1)

Generalizability theory permits analysis of different variance components that contribute to the variance in the *universe scores*. These variance components indicate the magnitude of error in generalizing from an object's (*object of measurement*) score on a single condition of each facet to its *universe score*. In addition to this, a *G* (generalizability) coefficient or a ϕ coefficient (index of dependability) (Brennan & Kane, 1977; Brennan, 2001) can also be calculated, both of which are analogous to the classical test theory reliability coefficient. In situations where absolute standings of objects of measurement are of concern, an index of dependability (ϕ coefficient) is calculated using the formula

$$\phi = \frac{\sigma^2_{(\tau)}}{\sigma^2_{(\tau)} + \sigma^2_{(\Delta)}},$$

where $\sigma^2_{(\tau)}$ is the variance attributable to the object of measurement and $\sigma^2_{(\Delta)}$ is the *absolute error variance*. The *absolute error variance* is defined as the variance of differences between an object's (*object of measurement*) observed and *universe score.* For the present study, absolute decisions are more relevant since it is important to reduce the error associated with using *observed score* on TALQ scales as *universe score* for a *class*.

Generalizability theory also permits predicting different variance components that would be expected if the number of conditions in the *facets* were changed. This would also help to redesign a possibly more efficient measurement.

Research Question 1: The two sources of error variance, when making reliability judgments from mean course evaluation scores for a class, are items and students. Therefore, the two facets included in the designs to answer this research question are *items* and *students*. The

object of measurement is the *class*. For analysis purposes, only one class taught by each instructor was included which resulted in 8 different classes taught by 8 different instructors.

The students *facet* (*s*) is nested within class (*c*), the *object of measurement*. The items *facet* (*i*) is crossed with both students and instructors. This is a partially nested design [(*s*:*c*) × *i*]. Figure 1 represents the variance components in a Venn diagram.

--Insert Figure 1--

The formula for calculating the index of dependability (ϕ) is

$$\phi = \frac{\sigma^2_{(\tau)}}{\sigma^2_{(\tau)} + \sigma^2_{(\Delta)}},$$

where $\sigma^2_{(\Delta)}$ is the *absolute error variance* and is calculated using the following formula

$$\sigma^2_{(\Delta)} = \sigma^2(s{:}\,c) + \sigma^2(i) + \sigma^2(ci) + \sigma^2(si{:}\,c, e).$$

$\sigma^2_{(\tau)}$ is equal to $\sigma^2_{(c)}$, which is the variance associated with the *object of measurement* (class).

The index of dependability obtained in this manner is for a class score by one student on one item. Estimates of ϕ coefficient values can also be obtained when generalizing over for different configurations of number of *students* and *items*. The *absolute error variance* for different *D* (decision) study configurations is calculated using the formula

$$\sigma^2_{(\Delta)} = \frac{\sigma^2(s{:}\,c)}{n'_s} + \frac{\sigma^2(i)}{n'_i} + \frac{\sigma^2(ci)}{n'_i} + \frac{\sigma^2(si{:}\,c, e)}{n'_i n'_s},$$

where n'_s is the number of students and n'_i is the number of items for the D study.
When the ϕ coefficient is calculated using this *absolute error*, it yields the dependability of scores for mean over n'_s students and n'_i items for a *class*.

Research Question 2: ALT and learning progress scale items attempt to measure student-level constructs. In case of these two TALQ scales, calculating mean ratings for a class over students is not appropriate. An instructor is interested in the dependability of mean student

ratings on items belonging to these scales. Feedback on mean scale scores for each student in a class would inform an instructor about the level of successful engagement in learning activities (ALT) and learning progress distribution. In this design the *absolute error variance* is calculated using the following formula

$$\sigma^2_{(\Delta)} = \sigma^2(i) + \sigma^2(ci) + \sigma^2(si{:}\,c, e).$$

$\sigma^2_{(\tau)}$, which is the variance associated with the *object of measurement*, is defined as

$\sigma^2_{(\tau)} = \sigma^2_{(c)} + \sigma^2(s{:}\,c)$ (Brennan, 2001).

Data Preparation: The numbers of participating students from the 12 classes were 44, 104, 16, 29, 22, 22, 49, 22, 26, 53, 35, and 42. If this entire data set were included in the estimation of variance components and ϕ coefficient, an unbalanced design would need to be used. In an unbalanced design, the number of conditions of at least one facet is not equal. For example, if different number of students (conditions of students facet) were used from each class, then the design would have been unbalanced. For an unbalanced design, different quadratic forms (Minimum Norm Quadratic Unbiased Estimation, Restricted Maximum Likelihood, Analysis of Variance) do not lead to the same estimates of variance components even though they are unbiased (Brennan, 2001). Therefore, the problem with unbalanced designs is that there are many estimators and no statistical basis for choosing among them. Moreover, statistical properties of variance estimates from diverse procedures have not yet been studied extensively (Brennan, 2001). On the other hand, in a balanced design, many quadratic forms lead to the same estimates of variance components. Therefore, a balanced design was used for the present study. Eliminating data to use a balanced design is a common practice in the literature and is preferred by many scholars (Shavelson & Webb, 1991).

To achieve a balanced design from such an unbalanced original data set requires random elimination of student data from each class until there are same numbers of students in each class. This would result in considerable loss of data for some classes and the discarded data may not be representative. Using all of the original data leads to more stable variance estimates. However, if the random sample data for each class is representative of the entire class, then using a balanced design by randomly eliminating data is appropriate.

In the present study, the smallest class in terms of the number of students without any missing data on any item was 15 for class 6. Therefore, for each class, 15 students were randomly selected from the students who responded to all 40 items. Differences between sample and entire class means on all 40 items for each class sample were compared to assess the representativeness of the sample. In general, the means for samples and entire class data were similar across classes for all items. This confirmed that the random samples chosen to be included in the generalizability study designs were more or less representative of the entire class data.

5. Findings

Research Question 1. Are mean student ratings for a semester-long *class* dependable, as measured by TALQ course evaluation scales established *a priori*?

Findings: Five out the seven TALQ scales, would yield close to dependable scores (Tables 1 to 7). Table 10 reports the dependability of scores for the TALQ scale scores along with the required number of items and students. Recommendations related to each scale are also reported.

Global instructor and course quality, student satisfaction, activation, and integration scales would yield dependable scores for a class when 15 students in a class rate 3 items on each

of these scales. In other words, for a class, these scale scores would be close to the *universe score* when averaging over the number of items (n_i') and the number of students (n_s') specified in table 10. The *universe score* is the scale score that a class would receive when all possible students in a class rate all possible items on the scale.

Four items were included in the student satisfaction scale on the TALQ instrument in this study. It is recommended that any 3 items be used since the scale yields dependable scores with 3 items and 15 students in a class. Even 2 items could be used in the case where time to complete the course evaluation is an issue, since the satisfaction scale would yield dependable scores even with 2 items (ϕ = .818 with 15 students). Items 6 (I am dissatisfied with this course) and 40 (I am very satisfied with this course) are similar except that item 6 was worded negatively. Therefore, it would be prudent to randomly select 1 out of these 2 items to be included on the TALQ course evaluation.

Following the same rationale, it is recommended to use 3 out of the 5 items on the activation scale that were included in this study. Considering that items 19 (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), 35 (In this course I was able to connect my past experience to new ideas and skills I was learning), and 36 (In this course I was *not* able to draw upon my past experience nor relate it to new things I was learning) are more similar to each other than to other items on the scale, it would be prudent to randomly select 1 out of these 3 items to be included on the TALQ course evaluation.

The activation scale would yield *highly* dependable scores (ϕ = .875) with 25 students in a class rating 5 items on the scale. On the other hand, the authentic problems (ϕ = .724) and the demonstration (ϕ = .674 with 5 items and 25 students) scales would not yield desirably

dependable scores even when 25 students in a class rate 5 items on each of these scales. Recommendations regarding these scales are discussed later in Section 6.

Research Question 2. Are mean ratings for *a student* dependable, based on self-reported academic learning time and learning progress scales?

The academic learning time scale would not yield a *highly* dependable score for a student even when averaging over 5 items ($\phi = 0.736$) (Table 7). Recommendations to improve the dependability of ALT scale are discussed below in section 6.

In contrast, in case of the learning progress scale, the score for a student would be dependable with only 2 items ($\phi = 0.887$) (Table 8). In other words, for a student, averaging over 2 items on the learning progress scale would yield a score close to the *universe score* on that scale. Therefore, it is recommended that any 2 out of the 5 items used in this study be used in on the learning progress scale. Considering that items 23 (I learned very little in this course) and 28 (I did not learn much as a result of taking this course) are opposite of item 10 (I learned a lot in this course) it would be prudent not to include more than 1 out of these 3 items.

6. Discussion

Potential problems with the TALQ course evaluation items and scales are discussed in this section. Some alternatives are also proposed to address the above-mentioned problems.

Authentic Problems Scale

The authentic problems scale was comprised of the following items:

(3) I performed a series of increasingly complex authentic tasks in this course.

(17) My instructor directly compared problems or tasks that we did, so that I could see how they were similar or different.

(22) I solved authentic problems or completed authentic tasks in this course.

(27) In this course I solved a variety of authentic problems that were organized from simple to complex.

(29) Assignments, tasks, or problems I did in this course are clearly relevant to my professional goals or field of work.

The authentic problems scale would yield only close to dependable scores even when averaging over 5 items and 25 students in a class ($\phi = 0.724$). The variance component for the students was 2.3 times than the one for the classes. In light of this finding, it is important to consider the within-class agreement for item 29 (Assignments, tasks, or problems I did in this course are clearly relevant to my professional goals or field of work). It was low in classes 1, 4, and 12 and was close to acceptable levels in class 3. Probably, these account for a relatively large variance component for the students nested within classes.

Another plausible explanation for low dependability of scale scores is that probably students responded differently to the five items on the authentic problems scale. This may be explaining the large variance component for confounded interaction terms. It is important to note that only items 29 and 17 (My instructor directly compared problems or tasks that we did, so that I could see how they were similar or different) on the scale did not include the term 'authentic'. All other items used the term 'authentic'. Potential problems with item 29 were discussed above. It is recommended that item 29 be excluded from the authentic problems scale because of the confounding of the relevance of tasks to a student's professional goals and instructor selection of specific tasks. When variance components were estimated after excluding item 29 from analysis, it yielded dependable or close to dependable scale scores ($\phi = 0.776$ with 4 items and 20 students per class; $\phi = 0.799$ with 4 items and 25 students per class).

Another reason for the low dependability of authentic problems scale scores is that probably the classes did not differ much in the use of authentic problems. It is noteworthy to contrast the results in the case of the authentic problems scale with the results in the case of the activation scale. It was found that the activation scale scores would be dependable ($\phi = 0.794$) when generalized over 3 items and 15 students. The relative variance components for confounded effects and for students are similar in both authentic problems and activation scales. However, for the activation scale, there was much larger variability among classes. The standard deviation for class mean scores was 0.428. Whereas, in the case of the authentic problems scale the variability among classes was low (SD=0.283). This explains the relatively higher ϕ coefficient values for activation scale as compared to those for the authentic problems scale. Nevertheless, the potential problems with authentic problem scales should not be overlooked.

A larger issue with the use of authentic problems scale on TALQ instrument could be that students were confused about the term 'authentic'. It was noted earlier that during the development process of TALQ, Frick et al. (2009) reported that faculty members who reviewed the course evaluation instrument pointed out confusion regarding the use of the term 'real-world'. Therefore, the term 'real-world' was changed to 'authentic' and following explanation was stated in the TALQ course evaluation:

"Note: In the items below, authentic problems or authentic tasks are 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.)."

Furthermore, the faculty members highlighted a potential problem that students may still attach different meanings to the term 'authentic problems'.

It is plausible that some students were confused about the relevance of tasks or problems to their professional field of work, or they did not have a chosen field of work at the time of taking the course. For example, in case of the senior level nursing class (class 11), students agreed with each other on all items belonging to the authentic problems scale. Students enrolled in this class had a chosen profession that they wished to pursue. Moreover, nursing students work with actual patients in the senior year of the academic program. Therefore, they may be in a better position to judge instructor's use of authentic problems as they are aware of the authentic problems. On the other hand, in the intermediate-level courses such as computer science (class 12) and history (class 3), the students did not agree as much with each other on the instructor's use of authentic problems. Students enrolled in these classes may not have a chosen profession at the time of taking these classes and therefore may not be able to assess the authenticity of the tasks and problems covered in class. Even if they had a chosen profession at this time, it is plausible they were not aware of the authentic problems.

While it may be easy to conclude that the measure of authentic problems should be excluded from the TALQ course evaluation because of the problems stated above, the present researchers argue that the measure of authentic problems is important as a measure of teaching effectiveness.

The importance of authentic problems in college classrooms was highlighted by a large-scale survey study of 81,499 high school students across 26 states in the U.S. (Yazzie-Mintz, 2007). Yazzie-Mintz (2007) reported that 66% of the students reported getting bored in class every day. The most frequent reason reported for the boredom was that the learning materials used in classes were irrelevant and uninteresting. Many students who considered dropping out of school reported that they did not see any value in the work done at school. The problems

reported in the results of this survey research are clearly related to the lack of authentic problems in school classes.

While the Yazzie-Mintz (2007) study uncovered problems with high school classes, the Commission on the Future of Higher Education (2006) report highlighted similar problems with college classes. The Commission reported that a frequent issue raised by employers with respect to college graduates is that many new graduates they hire are not well prepared to work. This lack of preparation of college graduates could be due to the lack of use of authentic problems in college classes. Considering the issues raised by the Commission, it becomes even more important to promote the use of authentic problems in college classes. Consequently, it is important to measure instructor's use of authentic problems since the feedback from a measure of the use of authentic problems would help instructors improve their courses.

Demonstration Scale

The demonstration scale was comprised of the following items:

(5) My instructor demonstrated skills I was expected to learn in this course.

(14) Media used in this course (texts, illustrations, graphics, audio, video, computers) were helpful in learning.

(16) My instructor gave examples and counter-examples of concepts that I was expected to learn.

(31) My instructor did not demonstrate skills I was expected to learn.

(38) My instructor provided alternative ways of understanding the same ideas or skills.

The dependability of the demonstration scale scores was found to be problematic (ϕ = 0.674 with 5 items and 25 students). Even when generalizing over 5 items and 50 students, this scale would not yield highly dependable scores (ϕ = 0.692).When comparing the respective variance component estimates across TALQ scales, it is important to note that in the case of the

demonstration scale, the variance attributable to classes was low. The classes did not differ much (SD=0.27) in the use of the demonstration principle. On the other hand, the classes differed relatively more in the case of the activation principle scale score (SD=0.428). Other variance components are similar in the case of both the scales. The activation scale would yield dependable scores ($\phi = 0.794$) when generalizing over 3 items and 15 students. In contrast, the demonstration scale would not yield dependable scores ($\phi = 0.674$) even when generalizing over 5 items and 25 students. Considering that the within-class agreement was generally acceptable for the demonstration scale items, one plausible explanation for low dependability is probably the low variation among classes.

A likely reason for low variability among classes on the demonstration scale scores is that the demonstration principle is likely to be included in most university classes—i.e., instructors typically lecture during class and illustrate ideas through modeling. The items on the demonstration scale are related to use of examples, presentation of skills or knowledge to be learned, media used in the course, and ways of understanding the concepts. It is reasonable to expect the presence of these aspects in a semester-long class in a wide variety of subject areas. For example, in the current study, in the history class (where generally the mean scale scores were low as compared to other classes) the mean score (M=3.69) on the demonstration scale was the second highest among the TALQ scales within the same class. Similarly, in the case of the computer science class (where generally the mean scale scores were low as compared to other classes) the mean scale score for demonstration (M=3.06) was the third highest among TALQ scales within the same class. Amongst the 8 classes included in this design, the lowest mean score on demonstration scale was 3.06. The variability between mean scores for classes on other scales was generally higher than that in the case of the demonstration scale.

The present researchers recommend that items on the demonstration scale should be included on the TALQ course evaluation since it would help identify instructors who do not use the demonstration principle.

Application Scale

The application scale was comprised of the following items:

(7) My instructor detected and corrected errors I was making when solving problems, doing learning tasks or completing assignments.

(32) I had opportunities to practice or try out what I learned in this course.

(37) My course instructor gave me personal feedback or appropriate coaching on what I was trying to learn.

The application scale would yield close to dependable scores ($\phi = 0.753$) when 25 students rate 3 items in a class. The dependability of the application scale scores ($\phi = 0.794$) would increase when 25 students rate on 5 items of the scale. Increasing the number of items would result in more time required to complete the TALQ course evaluation. Therefore, it is worthwhile to investigate for the presence of any problematic items in order to increase the dependability without increasing the number of items comprising the scale.

As was discussed earlier, the wording of item 37 (My course instructor gave me personal feedback or appropriate coaching on what I was trying to learn) may be problematic which is probably contributing to low dependability of the scale scores. Variance components were estimated without item 37. It resulted in higher dependability of scale scores ($\phi = 0.726$ Vs 0.686, with 3 items and 15 students; $\phi = 0.794$ Vs 0.725, with 5 items and 15 students) in general. It is important to note that after excluding item 37 from the scale there are only 2 items

remaining. Therefore, it is recommended that for future studies item 37 be revised. As discussed earlier, the revised item 37 is as follows:

My instructor gave me feedback on what I was trying to learn.

Integration

The integration scale comprised of the following items:

(11) I had opportunities in this course to explore how I could personally use what I have learned.

(24) I see how I can apply what I learned in this course to real life situations.

(30) I was able to publicly demonstrate to others what I learned in this course.

(33) In this course I was able to reflect on, discuss with others, and defend what I learned.

(39) I do not expect to apply what I learned in this course to my chosen profession or field of work.

Although the dependability of class mean scores on the integration scale was high (ϕ = 0.828 with 3 items and 15 students), low within-class agreement on item 39 was observed which should not be overlooked since it means that the students within a class did not agree with each other on this instructor-level variable. The wording of item 39 (I do not expect to apply what I learned in this course to my chosen profession or field of work) is problematic because of possible confounding between relevance to a student‘s professional goals and instructor selection of tasks. Therefore, it is recommended that item 39 be excluded from this scale. Five items on the integration scale were included in this study. After excluding item 39, it is recommended to use any 3 of the remaining 4 items.

Academic Learning Time Scale (Research Question 3)

The academic learning time comprised of the following items:

(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.

(13) I spent a lot of time doing tasks, projects and/or assignments, and my instructor judged my work as high quality.

(21) 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.

(25) I did a minimum amount of work and made little effort in this course.

The dependability of the ALT scale score for a student was investigated using a one-facet design [(s:c) × i] with students nested within classes and items crossed with both students and classes. The object of measurement was students nested within classes. Therefore, the variability due to students and due to classes both contributed to the variability due to the object of measurement.

In case of the ALT scale, the mean student scores on items were not highly dependable ($\phi = 0.736$ with 5 items). One recommendation is to use at least 5 items on the ALT scale in order to get close to dependable scale scores. Other possibilities could be explored to get dependable scores on the ALT scale while reducing the number of items.

Considering a large variance component for the confounded interaction and error effect (51.28% of the total variance), a plausible explanation for low dependability is that probably students responded differently to the five items on ALT scale. It is possible that the students may have gotten confused with the double stems in items 13 (I spent a lot of time doing tasks, projects and/or assignments, and my instructor judged my work as high quality), 21 (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), and 25 (I did a minimum amount of work and made little effort in this course). Whereas, for items 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) and 12 (I frequently did very good work on projects, assignments, problems and/or learning activities for this course) there was only one stem. Moreover, items 1 and 12 are related to student success on activities, whereas items 13 and 21 include the amount of time students spent doing academic tasks in addition to their success on these tasks. It is recommended that the following items on ALT scale be used in future studies:

(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.

(21) I put a great deal of effort into this course.

(12) I frequently did very good work on projects, assignments, problems and/or learning activities for this course

(13) I spent a lot of time doing tasks, projects and/or assignments.

Revised item 1 inquires about student success on tasks and revised item 21 is related to the amount of student effort. A high (strongly agree) response on the revised item 21 and a low (strongly disagree) response on the revised item 1 together would reflect more ALT. Similarly, a high response on both the revised item 12 and the revised item 13 would reflect more ALT. The responses to these items could be combined at the time of aggregation.

However, this more complex way of constructing the ALT scale may not be practical within many university course evaluation systems. While compound items should generally be avoided, ALT is nonetheless a compound concept—it requires student engagement in course tasks, and that such engagement is frequently successful. Hence, it may be more practical to

have more ALT items, rather than introducing an unusual way of summarizing course evaluation results. In other words, while fewer items might take students a few seconds less to complete, the cost of more items with respect to time to complete the survey is minimal. Most students completed the TALQ instrument in less than 10 minutes.

7. **Recommended TALQ scale items**

Scale	Items	Notes/Comments
Global Instructor and Course Quality Scale	8. Overall, I would rate the quality of this course as outstanding. 15. Overall, I would rate this instructor as outstanding. 34. Overall, I would recommend this instructor to others	
Student Satisfaction Scale	2. I am very satisfied with how my instructor taught this class. 6. I am dissatisfied with this course. 18. This course was a waste of time and money. 40. I am very satisfied with this course.	Use 2 or 3 of the original 4 items. It is suggested to use only one item between items 6 and 40 since they are similar except that they are worded negatively.
Academic Learning Time Scale	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. 21. I put a great deal of effort into this course. 12. I frequently did very good work on projects, assignments, problems and/or learning activities for this course 13. I spent a lot of time doing tasks, projects and/or assignments.	Combine the results of items 1 and 21 and items 12 and 13; or retain the original compound items and add several more items in order to keep scale scoring simpler.
Learning Progress Scale	4. Compared to what I knew before I took this course, I learned a lot. 10. I learned a lot in this course. 20. Looking back to when this course began, I have made a big improvement in my skills and knowledge in this subject. 23. I learned very little in this course. 28. I did not learn much as a result of taking this course.	Use 2 out of these 5 items. It is suggested not to include more than one item among items 10, 23, and 28 since they are similar to each other.

Authentic Problems Scale	3. I was expected to perform a series of increasingly complex authentic problems in this course. 17. My instructor directly compared problems or tasks that we did, so that I could see how they were similar or different. 22. I was expected to solve authentic problems or to complete authentic tasks in this course. 27. In this course I was expected to solve a variety of authentic problems that were organized from simple to complex.	
Activation Scale	9. I engaged in experiences that subsequently helped me learn ideas or skills that were new and unfamiliar to me. 19. 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. 26. My instructor provided a learning structure that helped me to mentally organize new knowledge and skills. 35. In this course I was able to connect my past experience to new ideas and skills I was learning. 36. In this course I was not able to draw upon my past experience nor relate it to new things I was learning.	Use 3 items on this scale. Considering that items 19, 35, and 36 are more similar to each other than the other items on the scale, it is suggested to use one of these three items.
Demonstration Scale	5. My instructor demonstrated skills I was expected to learn in this course. 14. Media used in this course (texts, illustrations, graphics, audio, video, computers) were helpful in learning. 16. My instructor gave examples and counter-examples of concepts that I was expected to learn. 31. My instructor did not demonstrate skills I was expected to learn. 38. My instructor provided alternative ways of understanding the same ideas or skills.	
Application Scale	7. My instructor detected and corrected errors I was making when solving problems, doing learning tasks or completing assignments. 32. I had opportunities to practice or try out what I learned in this course. 37. My instructor gave me feedback on what I was trying to learn.	

Integration Scale	11. I had opportunities in this course to explore how I could personally use what I have learned. 24. I see how I can apply what I learned in this course to real life situations. 30. I was able to publicly demonstrate to others what I learned in this course. 33. In this course I was able to reflect on, discuss with others, and defend what I learned.	Use any 3 items.

8. Significance

Teaching and Learning Quality course evaluation scales developed by Frick et al. (2009) include items from the *First Principles of Instruction* (Merrill, 2002; 2007; Merrill et al., 2009) and academic learning time (Rangel & Berliner, 2007). *First Principles of Instruction* were synthesized from instructional theories and models in the literature. Merrill (2002; 2007) argues that these principles are applicable to a wide variety of subject areas. These principles have also been related empirically to student learning achievement in a wide variety of course content areas. Similarly, academic learning time has been related empirically to student learning achievement. Feedback on items related to the *First Principles of Instruction* and academic learning time has the potential to inform what aspects of teaching need improvement. However, this was not investigated in the present study. Future research in this area, as discussed above, is warranted.

The usefulness of the feedback from course evaluations aligns very well with the objective of overall accountability movement in higher education in the United States which emphasizes improvement in the quality of teaching in postsecondary education.

References

Astin, A. (1991). *Assessment for excellence.* New York: American Council on Education and Macmillan Publishing Company.

Astin, A. (1993). *What matters in college? Four critical years revisited*. San Francisco: Jossey-Bass.

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.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: The American College Testing Program.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Brennan, R. L. & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14,* 277-289.

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

Ellis, E. S., Worthington, L. A. (1994). *Research synthesis on effective teaching principles and the design of quality tools for educators*. Eugene, Oregon: National Center to Improve the Tools for Educators.

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

Finn, J. D. (1989). Withdrawing from school. *Review of Educational Research, 59*(2), 117-142.

Finn, J. D., & Rock, D. A. (1997). Academic success among students at risk for school failure. *Journal of Applied Psychology, 82*, 221-234.

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.

Foshay, W. R. R., Silber, K. H., & Stelnicki, M. B. (2003). *Writing Training Materials That Work: How to Train Anyone to Do Anything*. San Francisco, CA: Pfeiffer.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School Engagement: Potential of the concept, state of the evidence. Review of Educational Research, 74(1), 59-109.

Frick, T.W., Chadha, R., Watson, C., Wang, Y., & Green, P. (2009). College student perceptions of teaching and learning quality. *Educational Technology Research and Development*, *57* (5), 705-720.

Frick, T.W., Chadha, R., Watson, C., & Zlatkovska, E. (2010). Improving course evaluations to improve instruction and complex learning in higher education. *Educational Technology Research and Development*, *58* (2), 115-136.

Friedman, M. I., & Fisher, S. P. (1998). *Handbook on effective instructional strategies*. Columbia, South Carolina: The Institute for Evidence-Based Decision-Making in Education.

Gagne, R. M. (1965). *The Conditions of Learning*. New York: Holt, Rinehart and Winston.

Gagne, R. M. (1985). *The Conditions of Learning and Theory of Instruction*, 4th ed. New York: Holt, Rinehart and Winston.

Gagne, R. M., Briggs L. J., & Wager, W. (1988). *Principles of Instructional Design.* Holt, Rinehart & Winston: New York.

Gagne, R. M., & Briggs L. J. (1974). *Principles of instructional design.* New York: Holt, Rinehart & Winston.

Gagne, R. M., Wager, W. W., Golas, K., Keller, J.M. (2004). *Principles of Instructional Design.* Wadsworth Publishing: Belmont, CA .

Glaser, R. (1992). *Advances in Instructional Psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jonassen, D. (1999). Designing constructivist learning environments. In C. M. Reigeluth (Ed.), *Instructional-design theories and models: A new paradigm of instructional theory* (Vol. 2) (pp. 215-239). Mahwah, New Jersey: Lawrence Erlbaum

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

Kuh, G. D. (2001). Assessing what really matters to student learning: Inside the National Survey of Student Engagement. *Change, 33*(3), 10-17, 66.

Kuh, G. D. (2003). What we're learning about student engagement from NSSE. *Change, 35*(2), 24-32.

Kuh, G.D., Kinzie, J., Buckley, J.A., Bridges B.K., & Hayek J.C. (2006). What matters to student success: A review of the literature. Retrieved January 15, 2008 from http://nces.ed.gov/IPEDS/research/pdf/Kuh_Team_Report.pdf.

L'Hommedieu, R., Menges, R.J., & Brinko, K.T. (1990). Methodological explanations for the modest effects of feedback from student ratings. *Journal of Educational Psychology, 82*(2), 232–241.

Lang, J.W. & Kersting, M. (2007). Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run? *Instructional Science, 35*, 187-205.

Marzano, R. J. (2003). What *works in schools: Translating research into action*. Alexandria, Virginia: Association for Supervision and Curriculum Development.

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement*. Alexandria, VA: Association for Supervision and Curriculum Development.

McCarthy, B. (1996). *About Learning*. Barrington, IL: Excel.

Merrill, M.D. (1994). *Instructional Design Theory*. Englewood Cliffs, NJ: Educational Technology Publications.

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

Merrill, M.D. (2007). First principles of instruction: a synthesis. In R.A. Reiser & j.V. Dempsey (Eds.), *Trends and Issues in Instructional Design and Technology*, 2nd ed. (Vol. 2. pp. 62-71). Upper Saddle River, NJ: Merrill/Prentice Hall.

Merril, M.D. (2009). *Instructional Design Theories and Models*, Vol. III. Mahwah, NJ: Lawrence Erlbaum Associates.

Merrill, M. D., Barclay, M., van Schaak, A. (2008). Prescriptive principles for instructional design. In M. J. Spector, J. van Merrienboer, & M. P. Driscoll (Eds.), *The handbook of*

research for educational communications and technology, 3rd edition, Routledge: New York.

Newmann, F. M., Wehlage, G. G., & Lamborn, S. D. (1992). The significance and sources of student engagement. In F. M. Newmann (Ed.), *Engagement and achievement in American secondary schools* (pp. 11-39). New York: Teachers College Press.

Rangel, E., & Berliner, D. (2007). Essential information for education policy: Time to learn. *Research Points: American Educational Research Association*, *5*(2), 1-4.

Reigeluth, C. M., Ed. (1983). *Instructional-Design Theories and Models: An Overview of Their Current Status*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Reigeluth, C. M., Ed. (1987). *Instructional Theories in Action: Lessons Illustrating Selected Theories and Models*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Reigeluth, C. M., Ed. (1999). *Instructional-Design Theories and Models: A New Paradigm of Instructional Theory*, Vol. II. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Schank, R.C., Berman, T.R. & Macperson, K.A. (1999). Learning by doing. In C.M. Reigeluth (Ed.), *Instructional design theories and models: A new paradigm of instructional theory* (Vol. II) (pp. 161–181). Mahwah, NJ: Lawrence Erlbaum Associates.

Schwartz, D., Lin, X., Brophy, S., & Bransford, J.D. (1999). Toward the development of flexibly adaptive instructional designs. In C.M. Reigeluth (Ed.), *Instructional design theories and models: A new paradigm of instructional theory* (Vol. II) (pp. 183–213). Mahwah, NJ: Lawrence Erlbaum Associates.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Smith, P. L. (1979). The generalizability of student ratings of courses: Asking the right questions. *Journal of Educational Measurement*, 16, 77-87.

Snyder, C. R., & Clair, M. (1976). Effects of expected and obtained grades on teacher evaluation and attribution of performance. *Journal of Educational Psychology*, *68*, 75-82.

Spencer, P. A. & Flyr, M. L. (1992). *The formal evaluation as an impetus to classroom change: Myth or reality?* (ERIC Document Reproduction Service No. ED349053).

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

Sugrue, B. (2004). Five instructional design principles worth revisiting. *The Criterion* (Spring), 10.

Tennyson, R. D., Schott, F., Seel, N. M., & Dijkstra, S., Eds. (1997). *Instructional Design International Perspective: Theory*, *Research*, *and Models*, Vol. 1. Mahwah, NJ: Lawrence Erlbaum Associates.

Thomson. (2002). *Thomson Job Impact Study: The Next Generation of Learning*. Naperville, IL: NETg (www.netg.com).

U.S. Department of Education. *A Test of Leadership: Charting the Future of American Higher Education*. Report of the Commission Appointed by Secretary of Education Margaret Spellings. Washington, D.C.: U.S. Department of Education, 2006.

van Merriënboer, J. J. G. (1997). *Training Complex Cognitive Skills: A Four-Component Instructional Design Model for Technical Training*. Englewood Cliffs, NJ: Educational Technology Publications.

Van Merriënboer, J. G., Clark, R.E., & de Croock, M, B. (2002). Blueprints for complex learning: The 4C/ID-Model. *Educational Technology Research and Development, 50* (2), 39-64.

van Merriënboer, J. J. G. and Kirschner, P. A. (2007). *Ten Steps to Complex Learning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Yazzie-Mintz, E. (2007). *Voices of students on engagement: A report on the 2006 high school survey of student engagement*. Retrieved April 15, 2009, from http://ceep.indiana.edu/hssse/pdf/HSSSE_2006_Report.pdf.

Table 1. G study (s:c) × i and D study (S:c) × I designs for the Global Instructor and Course Quality scale

Source	*Variance*	*Percentage of Total Variance*		*D studies*					
			n'_i	*3*	*3*	*3*	*5*	*5*	*5*
			n'_s	*15*	*20*	*25*	*15*	*20*	*25*
c	0.244	19.93%		0.244	0.244	0.244	0.244	0.244	0.244
s:c	0.647	52.86%		0.043	0.032	0.026	0.043	0.032	0.026
i	0.059	4.82%		0.020	0.020	0.020	0.012	0.012	0.012
ci	0.019	1.55%		0.006	0.006	0.006	0.004	0.004	0.004
si:c	0.255	20.83%		0.006	0.004	0.003	0.003	0.003	0.002
Total	1.224			0.319	0.307	0.299	0.306	0.295	0.288
$\sigma^2_{(\Delta)}$	0.98			0.075	0.063	0.055	0.062	0.051	0.044
ϕ	0.199			0.765	0.796	0.815	0.797	0.829	0.849

c – class, s – student, i – item, n'_i – number of items, n'_s – number of students, $\sigma^2_{(\Delta)}$ - absolute error variance, ϕ (Index of dependability)

Table 2. G study (s:c) × i and D study (S:c) × I designs for the Student Satisfaction scale

Source	*Variance*	*Percentage of Total Variance*		*D studies*					
			n'_i	*3*	*3*	*3*	*5*	*5*	*5*
			n'_s	*15*	*20*	*25*	*15*	*20*	*25*
c	0.322	25.95%		0.322	0.322	0.322	0.322	0.322	0.322
s:c	0.565	45.53%		0.038	0.028	0.023	0.038	0.028	0.023
i	0.043	3.46%		0.014	0.014	0.014	0.009	0.009	0.009
ci	0.005	0.40%		0.002	0.002	0.002	0.001	0.001	0.001
si:c	0.306	24.66%		0.007	0.005	0.004	0.004	0.003	0.002
Total	1.241			0.382	0.371	0.365	0.373	0.363	0.357
$\sigma^2_{(\Delta)}$	0.919			0.060	0.049	0.043	0.051	0.041	0.035
ϕ	0.259			0.842	0.867	0.883	0.862	0.887	0.903

c – class, s – student, i – item, n'_i – number of items, n'_s – number of students, $\sigma^2_{(\Delta)}$ - absolute error variance, ϕ (Index of dependability)

Table 3. G study (s:c) × i and D study (S:c) × I designs for the Authentic Problems scale

Source	*Variance*	*Percentage of Total Variance*	*D studies*					
			n'_i	3	3	3	5	5	5
			n'_s	15	20	25	15	20	25
c	0.08	9.10%		0.080	0.080	0.080	0.080	0.080	0.080
s:c	0.184	20.93%		0.012	0.009	0.007	0.012	0.009	0.007
i	0.003	0.34%		0.001	0.001	0.001	0.001	0.001	0.001
ci	0.092	10.47%		0.031	0.031	0.031	0.018	0.018	0.018
si:c	0.52	59.16%		0.012	0.009	0.007	0.007	0.005	0.004
Total	0.879			0.135	0.130	0.126	0.118	0.113	0.111
$\sigma^2_{(\Delta)}$	0.799			0.055	0.050	0.046	0.038	0.033	0.031
ϕ	0.091			0.590	0.618	0.635	0.677	0.705	0.724

c – class, s – student, i – item, n'_i – number of items, n'_s – number of students, $\sigma^2_{(\Delta)}$ - absolute error variance, ϕ (Index of dependability)

Table 4. G study (s:c) × i and D study (S:c) × I designs for the Activation scale

Source	*Variance*	*Percentage of Total Variance*	*D studies*					
			n'_i	*3*	*3*	*3*	*5*	*5*	*5*
			n'_s	*15*	*25*	*20*	*15*	*20*	*25*
c	0.183	18.98%		0.183	0.183	0.183	0.183	0.183	0.183
s:c	0.284	29.46%		0.019	0.011	0.014	0.019	0.014	0.011
i	0.017	1.76%		0.006	0.006	0.006	0.003	0.003	0.003
ci	0.039	4.05%		0.013	0.013	0.013	0.008	0.008	0.008
si:c	0.441	45.75%		0.010	0.006	0.007	0.006	0.004	0.004
Total	0.964			0.230	0.219	0.223	0.219	0.213	0.209
$\sigma^2_{(\Delta)}$	0.781			0.047	0.036	0.040	0.036	0.030	0.026
ϕ	0.190			0.794	0.836	0.820	0.836	0.860	0.875

c – class, s – student, i – item, n'_i – number of items, n'_s – number of students, $\sigma^2_{(\Delta)}$ - absolute error variance, ϕ (Index of dependability)

Table 5. G study (s:c) × i and D study (S:c) × I designs for the Demonstration scale

Source	*Variance*	*Percentage of Total Variance*	*D studies*					
			n'_i *3*	*3*	*3*	*5*	*5*	*5*
			n'_s *15*	*20*	*25*	*15*	*20*	*25*
c	0.073	8.34%	0.073	0.073	0.073	0.073	0.073	0.073
s:c	0.295	33.71%	0.020	0.015	0.012	0.020	0.015	0.012
i	0.079	9.03%	0.026	0.026	0.026	0.016	0.016	0.016
ci	0.022	2.51%	0.007	0.007	0.007	0.004	0.004	0.004
si:c	0.406	46.40%	0.009	0.007	0.005	0.005	0.004	0.003
Total	0.875		0.135	0.128	0.124	0.118	0.112	0.108
$\sigma^2_{(\Delta)}$	0.802		0.062	0.055	0.051	0.045	0.039	0.035
ϕ	0.083		0.539	0.569	0.589	0.617	0.652	0.674

c – class, s – student, i – item, n'_i – number of items, n'_s – number of students, $\sigma^2_{(\Delta)}$ - absolute error variance, ϕ (Index of dependability)

Table 6. G study (s:c) × i and D study (S:c) × I designs for the Application scale

Source	*Variance*	*Percentage of Total Variance*	*D studies*					
			n'_i *3*	*3*	*3*	*5*	*5*	*5*
			n'_s *15*	*20*	*25*	*15*	*20*	*25*
c	0.15	12.24%	0.150	0.150	0.150	0.150	0.150	0.150
s:c	0.587	47.92%	0.039	0.029	0.023	0.039	0.029	0.023
i	0.003	0.24%	0.001	0.001	0.001	0.001	0.001	0.001
ci	0.057	4.65%	0.019	0.019	0.019	0.011	0.011	0.011
si:c	0.428	34.94%	0.010	0.007	0.006	0.006	0.004	0.003
Total	1.225		0.219	0.206	0.199	0.207	0.196	0.189
$\sigma^2_{(\Delta)}$	1.075		0.069	0.056	0.049	0.057	0.046	0.039
ϕ	0.122		0.686	0.726	0.753	0.725	0.767	0.794

c – class, s – student, i – item, n'_i – number of items, n'_s – number of students, $\sigma^2_{(\Delta)}$ - absolute error variance, ϕ (Index of dependability)

Table 7. G study (s:c) × i and D study (S:c) × I designs for the Integration scale

Source	*Variance*	*Percentage of Total Variance*	*D studies*					
			n'_i	*3*	*3*	*3*	*5*	*5*	*5*
			n'_s	*15*	*20*	*25*	*15*	*20*	*25*
c	0.304	26.93%		0.304	0.304	0.304	0.304	0.304	0.304
s:c	0.255	22.59%		0.017	0.013	0.010	0.017	0.013	0.010
i	0	0.00%		0.000	0.000	0.000	0.000	0.000	0.000
ci	0.107	9.48%		0.036	0.036	0.036	0.021	0.021	0.021
si:c	0.463	41.01%		0.010	0.008	0.006	0.006	0.005	0.004
Total	1.129			0.367	0.360	0.356	0.349	0.343	0.339
$\sigma^2_{(\Delta)}$	0.825			0.063	0.056	0.052	0.045	0.039	0.035
ϕ	0.269			0.828	0.844	0.854	0.872	0.887	0.896

c – class, s – student, i – item, n'_i – number of items, n'_s – number of students, $\sigma^2_{(\Delta)}$ - absolute error variance, ϕ (Index of dependability)

Table 8. G study (s:c) × i and D study (s:c) × I designs for the ALT scale

Source	*Variance*	*Percentage of Total Variance*	*D studies*				
			n'_i	*2*	*3*	*4*	*5*
c	0.073	7.80%		0.073	0.073	0.073	0.073
s:c	0.262	27.99%		0.262	0.262	0.262	0.262
i	0.069	7.37%		0.035	0.023	0.017	0.014
ci	0.052	5.56%		0.026	0.017	0.013	0.010
si:c	0.48	51.28%		0.240	0.160	0.120	0.096
Total	0.936			0.636	0.535	0.485	0.455
$\sigma^2_{(\Delta)}$	0.601			0.301	0.200	0.150	0.120
ϕ	0.358			0.527	0.626	0.690	0.736

c – class, s – student, i – item, n'_i number of items, $\sigma^2_{(\Delta)}$ - absolute error variance, ϕ (Index of dependability)

Table 9. G study (s:c) × i and D study (s:c) × I designs for the Learning Progress scale

Source	*Variance*	*Percentage of Total Variance*	n'_i	*D studies* 2	3	4	5
c	0.195	20.59%		0.195	0.195	0.195	0.195
s:c	0.56	59.13%		0.560	0.560	0.560	0.560
i	0.013	1.37%		0.007	0.004	0.003	0.003
ci	0.004	0.42%		0.002	0.001	0.001	0.001
si:c	0.175	18.48%		0.088	0.058	0.044	0.035
Total	0.947			0.851	0.819	0.803	0.793
$\sigma^2_{(\Delta)}$	0.752			0.096	0.064	0.048	0.038
ϕ	0.206			0.887	0.922	0.940	0.952

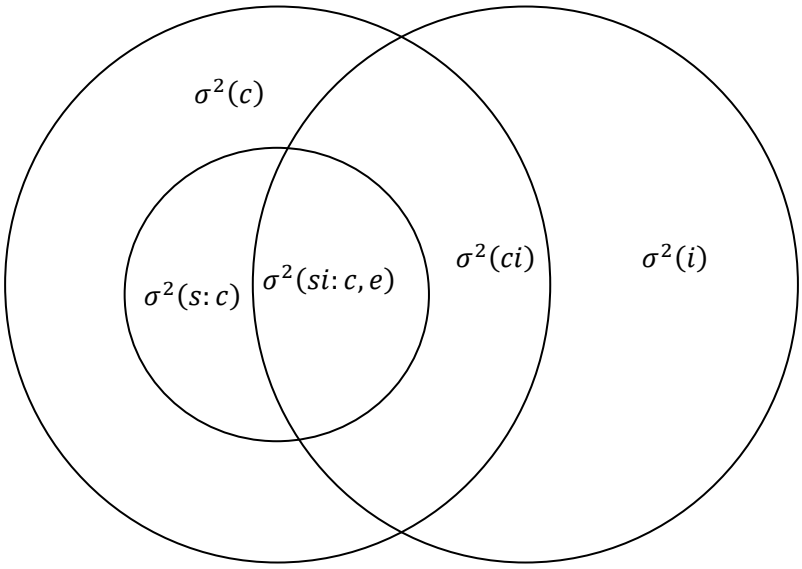
c – class, s – student, i – item, n'_i number of items, $\sigma^2_{(\Delta)}$ - absolute error variance, ϕ (Index of dependability)

Table 10. Dependability of the TALQ scale scores (summary of results for research question 1)

TALQ Scale	i	n_i'	n_s'	ϕ	Recommendations
Global Instructor and Course Quality	3	3	15	0.765	Use the original 3 items
Student Satisfaction	4	3	15	0.842	Use 3 of the original 4 items
Authentic Problems	5	5	25	0.724	Remove item 29 from this scale
Activation	5	3	15	0.794	Use 3 of the original 5 items
Demonstration	5	5	25	0.674	Include classes with variability in the use of the demonstration principle
Application	3	3	25	0.753	Modify item 37
Integration	5	3	15	0.828	Remove item 39 and use 3 of the remaining 4 items

i –number of items in the TALQ instrument, n'_i – number of items, n'_s – number of students, ϕ (Index of dependability)

Figure 1.Variance components for two-facet (s:c) × i design



Appendix A: Teaching and Learning Quality Course Evaluation Instrument

Course Evaluation

Directions: Please answer each question by circling your choice.

Your Gender Male Female

Please answer the following questions about this class:

a. I would rate this class as:

1 2 3 4 5 6 7 8 9 10
Really awful (poor) About average Really great (outstanding)

b. In this course, I expect to receive (or did receive) a grade of:

A
B
C
D
F
Not applicable (or don't know)

c. I am a:

Freshman
Sophomore
Junior
Senior
Graduate Student
Other

d. With respect to achievement of objectives of this course, I consider myself a:

1 2 3 4 5 6 7 8 9 10
Non-master Master

Note: In the items below, *authentic problems* or *authentic tasks* are 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.).

If there were multiple instructors for this course, please rate the group of instructors as a whole.

Please rate each item by circling below as: SA=strongly agree, A=agree, U=undecided, D=Disagree, SD=strongly disagree, or NA=not applicable. Please rate all items.

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.	SA	A	U	D	SD	NA
2.	I am very satisfied with how my instructor taught this class.	SA	A	U	D	SD	NA
3.	I performed a series of increasingly complex authentic tasks in this course.	SA	A	U	D	SD	NA
4.	Compared to what I knew before I took this course, I learned a lot.	SA	A	U	D	SD	NA
5.	My instructor demonstrated skills I was expected to learn in this course.	SA	A	U	D	SD	NA
6.	I am dissatisfied with this course.	SA	A	U	D	SD	NA
7.	My instructor detected and corrected errors I was making when solving problems, doing learning tasks or completing assignments.	SA	A	U	D	SD	NA
8.	Overall, I would rate the quality of this course as outstanding.	SA	A	U	D	SD	NA
9.	I engaged in experiences that subsequently helped me learn ideas or skills that were new and unfamiliar to me.	SA	A	U	D	SD	NA
10.	I learned a lot in this course.	SA	A	U	D	SD	NA
11.	I had opportunities in this course to explore how I could personally use what I have learned.	SA	A	U	D	SD	NA
12.	I frequently did very good work on projects, assignments, problems and/or learning activities for this course.	SA	A	U	D	SD	NA
13.	I spent a lot of time doing tasks, projects and/or assignments, and my instructor judged my work as high quality.	SA	A	U	D	SD	NA

14. Media used in this course (texts, illustrations, graphics, audio, video, computers) were helpful in learning.	SA	A	U	D	SD	NA
15. Overall, I would rate this instructor as outstanding.	SA	A	U	D	SD	NA
16. My instructor gave examples and counter-examples of concepts that I was expected to learn.	SA	A	U	D	SD	NA
17. My instructor directly compared problems or tasks that we did, so that I could see how they were similar or different.	SA	A	U	D	SD	NA
18. This course was a waste of time and money.	SA	A	U	D	SD	NA
19. 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.	SA	A	U	D	SD	NA
20. Looking back to when this course began, I have made a big improvement in my skills and knowledge in this subject.	SA	A	U	D	SD	NA
21. 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.	SA	A	U	D	SD	NA
22. I solved authentic problems or completed authentic tasks in this course.	SA	A	U	D	SD	NA
23. I learned very little in this course.	SA	A	U	D	SD	NA
24. I see how I can apply what I learned in this course to real life situations.	SA	A	U	D	SD	NA
25. I did a minimum amount of work and made little effort in this course.	SA	A	U	D	SD	NA
26. My instructor provided a learning structure that helped me to mentally organize new knowledge and skills.	SA	A	U	D	SD	NA
27. In this course I solved a variety of authentic problems that were organized from simple to complex.	SA	A	U	D	SD	NA
28. I did not learn much as a result of taking this course.	SA	A	U	D	SD	NA
29. Assignments, tasks, or problems I did in this course are clearly relevant to my professional goals or field of work.	SA	A	U	D	SD	NA
30. I was able to publicly demonstrate to others what I learned in this course.	SA	A	U	D	SD	NA

31. My instructor did not demonstrate skills I was expected to learn.	SA	A	U	D	SD	NA
32. I had opportunities to practice or try out what I learned in this course.	SA	A	U	D	SD	NA
33. In this course I was able to reflect on, discuss with others, and defend what I learned.	SA	A	U	D	SD	NA
34. Overall, I would recommend this instructor to others.	SA	A	U	D	SD	NA
35. In this course I was able to connect my past experience to new ideas and skills I was learning.	SA	A	U	D	SD	NA
36. In this course I was not able to draw upon my past experience nor relate it to new things I was learning.	SA	A	U	D	SD	NA
37. My course instructor gave me personal feedback or appropriate coaching on what I was trying to learn.	SA	A	U	D	SD	NA
38. My instructor provided alternative ways of understanding the same ideas or skills.	SA	A	U	D	SD	NA
39. I do not expect to apply what I learned in this course to my chosen profession or field of work.	SA	A	U	D	SD	NA
40. I am very satisfied with this course.	SA	A	U	D	SD	NA

PLEASE GIVE THIS TO THE **RESEARCHER.**

Appendix B: Items on the Original Nine TALQ Scales

Academic Learning Time (ALT) scale

(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.
(13) I spent a lot of time doing tasks, projects and/or assignments, and my instructor judged my work as high quality.
(21) 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.
(25) I did a minimum amount of work and made little effort in this course.

Learning Progress scale

(4) Compared to what I knew before I took this course, I learned a lot.
(10) I learned a lot in this course.
(20) Looking back to when this course began, I have made a big improvement in my skills and knowledge in this subject.
(23) I learned very little in this course.
(28) I did not learn much as a result of taking this course.

Student satisfaction scale

(2) I am very satisfied with how my instructor taught this class.
(6) I am dissatisfied with this course.
(18) This course was a waste of time and money.
(40) I am very satisfied with this course.

Global Course and Instructor Quality scale

(8) Overall, I would rate the quality of this course as outstanding.
(15) Overall, I would rate this instructor as outstanding.
(34) Overall, I would recommend this instructor to others.

Authentic Problems Scale

(3) I performed a series of increasingly complex authentic tasks in this course.
(17) My instructor directly compared problems or tasks that we did, so that I could see how they were similar or different.
(22) I solved authentic problems or completed authentic tasks in this course.
(27) In this course I solved a variety of authentic problems that were organized from simple to complex.
(29) Assignments, tasks, or problems I did in this course are clearly relevant to my professional goals or field of work.

Activation scale

(9) I engaged in experiences that subsequently helped me learn ideas or skills that were new and unfamiliar to me.
(19) 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.
(26) My instructor provided a learning structure that helped me to mentally organize new knowledge and skills.
(35) In this course I was able to connect my past experience to new ideas and skills I was learning.
(36) In this course I was not able to draw upon my past experience nor relate it to new things I was learning.

Demonstration scale

(5) My instructor demonstrated skills I was expected to learn in this course.
(14) Media used in this course (texts, illustrations, graphics, audio, video, computers) were helpful in learning.
(16) My instructor gave examples and counter-examples of concepts that I was expected to learn.
(31) My instructor did not demonstrate skills I was expected to learn.
(38) My instructor provided alternative ways of understanding the same ideas or skills.

Application scale

(7) My instructor detected and corrected errors I was making when solving problems, doing learning tasks or completing assignments.
(32) I had opportunities to practice or try out what I learned in this course.
(37) My course instructor gave me personal feedback or appropriate coaching on what I was trying to learn.

Integration scale

(11) I had opportunities in this course to explore how I could personally use what I have learned.
(24) I see how I can apply what I learned in this course to real life situations.
(30) I was able to publicly demonstrate to others what I learned in this course.
(33) In this course I was able to reflect on, discuss with others, and defend what I learned.
(39) I do not expect to apply what I learned in this course to my chosen profession or field of work.