A Confirmatory Factor Analysis of the Latent Structure and Measurement Invariance in the University of Utah's Student Course Feedback Instrument

Mark St. André

Master's in Statistics in Educational Psychology Project
University of Utah

April 22, 2012

Running Head: CFA of Latent Structure in Utah's SCF

Abstract

Confirmatory factor analysis is used to examine the latent constructs and measurement invariance of the University of Utah's 14-question Student Course Feedback (SCF) instrument. All of the student responses from 13 departments within 6 colleges (n=23,516) for the Spring 2009 semester were analyzed. A four-factor measurement model was determined to have better fit than both a one-factor and two-factor model. The four-factor model was shown to have configural, metric, and scalar invariance across type of college (qualitative vs. quantitative), instructor gender, and instructor ethnicity (white vs. foreign). Suggestions for ongoing studies of the SCF instrument's validity are made.

Introduction

Student evaluations of teaching (SET's) are almost omnipresent in higher education. When students finish taking a college class, and almost always before they see their grades, they are asked to evaluate the course and the instructor. Because they are so common, there is an abundance of research on the topic. There is even a substantial number of meta-analyses of the literature over the past 20 years or so (Gravestock & Gregor-Greenleaf, 2008; Cashin, 1995; Cashin, 1988; Centra, 1993; Braskamp & Ory, 1994). This has led Cashin (1995), in a summary of the meta-analyses to remark: "There are probably more studies of student ratings than of all the other data used to evaluate college teaching combined" (p. 1).

The purpose of SET's has become two-fold: to evaluate instruction, and to give feedback to students as the consumers of courses. In fact, SET's have become the most common method for evaluating instructors in post-secondary education (Gravestock & Gregor-Greenleaf, 2008; Merritt, 2008; Benton, 2011). As such, many schools also use student evaluations for faculty retention, promotion, and tenure (RPT) reviews (Seldin, 1985; in Paswan and Young, 2002; Marsh & Roche, 1992; Fischer, 2004; Benton, 2011) and they often carry the most weight in those reviews (Franklin, 2001). This makes SET's a high-stakes endeavor.

Because SET's are high-stakes, whether institutions design their own or use one of the existing SET instruments on the market, it is crucial that administrators regularly assess the validity of their instruments: that is, understand what exactly it is that the instruments are measuring. Regular validity studies provide administrators with the data

that they need to assert that SET's are measuring what they think they are, and the opportunity to adjust the instrument if it doesn't. This assessment process of studying the data and adjusting the instrument accordingly will increase the validity of the important faculty decisions being made with this information.

Background on SET's

In the early 1950's, a number of schools began regularly seeking input from students on instructor effectiveness. The University of Washington was one of the first (Guthrie, 1954; in Clayson, 2008). Around the 1990's, the use of SET's increased radically to the point where four out of five institutions were doing some sort of regular SET. In his 1995 meta-analysis of the field, Cashin found more than 1,500 references to research being done on SET's. At the present, it would be very rare for a student to take a class and not be asked in some way to provide feedback on their instructor, the course, or both. This feedback almost always takes the shape of some sort of Likert scale rating as well as open-ended feedback (Gravestock and Gregor-Greenleaf, 2008).

There are two kinds of SET's that are in use. Commercially-available instruments, such as the Student Evaluation of Educational Quality (SEEQ; Marsh, 1982), and the Student Ratings of Instruction (IDEA Center, 2012), are examples of commercially available instruments. These instruments are regularly studied and evaluated by the vendors. The other kind of SET in use is the homegrown product. The current paper focuses on some validity issues that are relevant to schools that are using their own instruments. However, it is also important for those who are using commercial products to study them in the context of their institutions.

The meta-analyses of Cashin (1995) and Gravestock & Gregor-Greenleaf (2008) have both indicated that, where they have been studied, SET's have been shown to largely be reliable and valid. This, however, is not comforting. It is very likely that many instruments that are being used in the field have not been studied. Others have been studied, but the results have not been published. And, because of the bias of journals to publish positive findings, it is, overall very hard to say how accurate Cashin (1995) and Gravestock & Gregor-Greenleaf's (2008) finding is. Most importantly, however, is the need for every instrument in use to be studied in the context in which it is being used. This is good measurement practice, but also particularly important because of the use of SET's as primary indicators of instructor effectiveness in employee reviews.

The research questions of the current paper focus largely on validity issues, but a brief review of both reliability and validity issues in SET's is presented to provide context and background.

Reliability

Reliability is the consistency with which an instrument measures something in the same way every time it is used. To use a simple example, a thermometer is reliable if, with an acceptable amount of error, it reads the same temperature every time you use it under the same atmospheric conditions. The reliability of SET's is typically operationalized as the consistency, stability, and generalizability of student responses. The consistency has been demonstrated by showing, for example, rating agreement from students in the same class (Marsh & Roche, 1997) as well as for students across multiple classes (Ory & Ryan, 2001). Stability has been shown in retrospective studies where students' initial ratings are

correlated with ones taken a year or so later (Feldman, 1989). Generalizability captures the confidence evaluators have that student ratings of teachers are consistent with ratings of those same teachers in other courses, and this has also been shown. For example, Marsh (1982) compared ratings of teachers who were teaching the same classes and different classes and compared them with different teachers teaching the same classes, and with different teachers teaching different classes. He found that the instructor, and not the course, was more highly correlated with the student ratings. Overall, the reliability of SET's does not seem to be in question (Hobson, 2001).

Validity

Validity is a statement about the confidence one has that an instrument is measuring what one thinks it is measuring. Using the same metaphor as above, a thermometer is valid if evidence can be provided that shows it is measuring the temperature of the air and doing it accurately. Validity is also an argument about the appropriateness of conclusions drawn from the use of one's instrument with a particular population at a particular point in time (Messick, 1988). As such, validity is never "established," except for its use at one of those times with a particular population and context. It is therefore not surprising that, while the meta-analyses of SET's indicate general agreement on their validity (again, see Cashin, 1995; and Gravestock & Gregor-Greenleaf; 2008) there is a large amount of equivocation in the findings.

One common way that validity, in general, is demonstrated is to show that an instrument produces results that are similar to other instruments or data measuring the same construct (convergent validity). If the assumption is that the primary purpose of

SET's is to measure the effectiveness of teaching, then SET validity studies should aim to show that SET's correlate positively with another indicator of effective teaching and not with factors that are unrelated to teaching (bias). The challenge here is that there is no commonly accepted definition of "effective teaching" (Clayson, 2009). In an attempt to represent effective teaching, studies in this area have compared SET's with a variety of indicators, all of which have their own validity challenges, including: student grades, instructor self-rating, instructor peer ratings, and students' self-perception of how much they have learned. Because they are the most relevant to the current paper, a discussion of the relationship of SET's and course grade, SET's and students' self-perception of how much they have learned, and bias is given below.

SET's and Course Grade: One regular criticism of course evaluations is that they can simply be explained by course grades: high grade = high course evaluation (Greenwald & Gilmore, 1997; Feldman, 1997). However, the argument can be made that these relationships can be better explained by including in the analysis other important factors such as how well the students were prepared, how hard they tried, and last, but not least, how much they learned during the course (Clayson, 2008). All three of these factors could mediate the relationship between grade and evaluation: students who are well prepared, try hard and learn a lot might be more satisfied with a course. In fact, Cashin (1998) showed that grade expected, student motivation for the topic, and the instructor's method of delivery all contributed significantly to explaining student ratings of instructors. Cashin (1995) also showed that, across classes, students give higher ratings in the courses where they worked the hardest. McKeachie (1997) believes that teachers tend to target the

difficulty level of a class to better students, and as a result those students who are bound to do well rate it well because the course was appropriate and challenging for them.

SET's and Self-Perception of Learning: Individual student grades are often not made available because it means asking students' permission to violate the anonymity that is usually promised to students when they are asked to complete SET's. Because of this, students' self-perception of learning, a commonly asked question on SET's, is often substituted as a means of testing the validity of the instrument.

Some authors believe students are, in fact, the best qualified to rate their own learning (Cruse, 1987; Machina, 1987; both cited in Clayson, 2009), but they are in the minority. Many authors consider this method of testing validity to be problematic because students' perceptions of what they know is, of course, biased or limited by what they do, in fact, know. Less knowledgeable students believe they know more than they do because they are unaware of their lack of knowledge and at the same time smarter students are aware of what they don't know and therefore rate themselves more poorly (Grimes, 2002). This effect, if accurate, would flatten the slope of the relationship between SET and self-perception. Nonetheless, because the question is often included in SET's, students' self-perception is often the only readily available data on teacher quality with which to examine the relationship between learning and SET.

Bias: A validity-related concern among many authors is the question of whether SET's have biases, which Merritt (2008) defines as student, teacher, or course characteristics that affect student ratings but which are unrelated to any criteria of good

teaching. Some examples of bias include the charisma effect of instructors (Shevlin et al, 2000), gender bias, and ethnicity bias (see Merritt, 2008 for a discussion of bias). Martin and Radmacher (2001) showed that only teacher's extraversion predicted student evaluations after controlling for course grades, student age, and enrollment status. In an often-cited study, Ambady & Rosenthal (1993) showed that instructor and course evaluation ratings of students who watched a videotape of the first 30 seconds of a class correlated .75 with those students who completed the entire course and then evaluated the instructor. These findings seem to indicate the possibility that SET's are not entirely based on content of the class or the ability of the professor to impart knowledge, but instead on something more external, such as likeability, outgoingness, nonverbal behaviors or other external characteristics. Of course, it is also possible that outgoingness in a professor might be associated with other characteristics that make for a more welcoming classroom atmosphere where students feel they are able to engage in the learning process.

Ethnicity Bias: The issue of ethnicity bias is, surprisingly, not widely addressed in the literature (Centra, 1993), a fact that, as Merritt (2008) points out, is itself quite troubling. Despite minority professors raising their concerns about the fairness of SET's and potential bias, universities do not seem willing to address it (Huston, 2006). This is quite surprising, given the fact that so many schools use numerical averages of SET questions in retention, promotion, and tenure reviews. Where it has been studied, research shows that minorities receive lower ratings than non-minorities. Delgado and Bell (2005: in Merritt, 2008) showed that at a large, southern university evaluations for minorities were significantly lower than for non-minorities even when controlling for the level of the

course and tenure level. Even more troubling was the observation by some minority faculty that in written comments students seemed to give them both positive and negative feedback at the same time for the same course that contradicted itself. They thought this indicated that there might be another factor at play in the evaluation other than how well the instructor delivered the material and the course.

Smith (2007) took the examination of this issue a step further. She noted that some of the minority faculty at her institution thought anecdotally that they did quite well on the "multidimensional" items (questions about specific skills) on their SET, but poorly on the two overall questions (one for instructor and one for course). She examined these phenomena in the data and found that white faculty received significantly higher multi-dimensional scores than blacks (other ethnicities were left out because of small sample size), and the same pattern was true for their scores on the overall items. However, she did not make an attempt to statistically control for multidimensional item scores in predicting global scores. This would have helped answer the question of whether black faculty received lower global ratings when controlling for the multidimensional (specific skill) ratings and might indicate some bias based on ethnicity.

***Gender Bias*:** The research is less clear on the existence of gender bias. Two studies by Feldman (1992, 1993) showed mostly inconclusive evidence of same gender or cross-gender bias in student ratings of instructors. Other studies have shown slight bias of male students against female faculty: Basow and Silberg (1987) identified 16 pairs of instructors (one male and one female) on their campus that were matched on rank, disciplinary area, and years of experience and found that male students gave female

instructors lower ratings than they gave to male instructors. Centra (2000) showed a significant bias of female students to rate female faculty more highly than male students do, but found no difference in the ratings of male instructors by male and female students. It is unclear whether these findings indicate a bias among female students for female faculty or a male bias against female faculty.

Females may learn better based on the ways in which female faculty teach compared to males. Female faculty are less likely to use lecture and more likely to use discussion groups and to value connection over separation of information (Belenky, Clinchy, Goldberger, and Tarule, 1986; cited in Centra, 2000). This style may resonate more with female students and less so with males, and would explain the difference in ratings by males and females of female instructors. Centra (2000) states that the effect of gender bias and the effect of the interaction of gender of student and instructor on ratings is inconclusive and warrants further study with better controls.

***Construct Validity*:** Cashin's (1995) meta-analysis says that there are six typical constructs that SET's are designed to measure:

1. Course organization and planning
2. Clarity, communication skills
3. Teacher and student interaction, rapport
4. Course difficulty, workload
5. Grading and examinations
6. Student self-rated learning (Cashin, 1995; p.2)

Gravestock and Gregorleaf (2008) suggest that all questions used in an SET instrument should be aligned with a well-developed theoretical construct. It is also common, however, to ask global questions about the instructor and the course as part of the SET. These global questions, however, do not typically fall into a specific construct. It is reasonable to think

about these two kinds of questions – construct-aligned, and global – along the lines of formative and summative evaluation. In the context of SET's, the purpose of construct-aligned items is to produce data for formative evaluation: providing feedback that improves teacher performance. The purpose of the overall questions is to provide summative evaluation: an overall assessment of the instructor and/or course quality.

Another way in which an instrument's validity can be demonstrated is to show that it is measuring the constructs it was intended to measure. Psychological instruments are all designed to measure some kind of construct – intelligence, anxiety, depression – that cannot be directly observed. While individual items are only approximations of the underlying construct that is trying to be measured, several similar items together can provide a valid indirect assessment of the attribute under study (Gregorich, 2006). The process of confirming that the intended constructs actually appear in the response data is a logical step in any analysis of an instrument's validity. It answers one of the core questions of validity: "Did we measure what we intended to measure with this instrument, or is the data telling us we measured something different?"

This analysis can be done through a variety of means, but the most common is confirmatory factor analysis. In this procedure, a matrix of correlations between all of the observed items is examined in at attempt to explain as much variance as possible between the items with a fewer number of unobserved variables, or factors. In his article on the importance of evaluating measurement invariance in self-report instruments, Gregorich (2006) describes the appropriateness of using confirmatory factor analysis (CFA) for evaluating construct validity and measurement invariance:

> The confirmatory factor analysis (CFA) framework provides a means to test the construct validity of item sets, i.e., whether item sets are indirect measures of

> hypothesized latent variables. Furthermore, CFA can test whether evidence of construct validity is invariant across 2 or more population groups…The results of these tests help to determine which types of quantitative group comparisons are defensible (p. S79, Gregorich, 2006).

Gregorich's focus on the appropriateness of making comparisons between groups based on the instrument's demonstrated measurement invariance is an important and relevant one for the study of SET's. One key element of validity that has not yet been discussed is *consequential validity* – what the consequences are regarding the use of the instrument under study. If it is known that the instrument will be used for particular purposes, then it is important for those studying the instrument to consider that in the validity analysis. This point is particularly relevant for SET's, given the evidence that they are the most commonly used metric for demonstrating teacher effectiveness, and are surely used to make comparisons among teachers and courses. If the construct validity evidence that is available cannot be shown to be invariant across salient population groups, it might be inappropriate to make comparisons between those groups with the instrument.

Summary of Validity in SET's: In 1997, American Psychologist released a special edition on course evaluations, in which the authors conclude that SET's are the "single most valid source on teaching effectiveness" (McKeachie, 1997, p. 1218). But it might be more accurate to say that SET's are the most valid source on teaching effectiveness *that is available*. Many believe that statement to be a more accurate assessment of the state of affairs in this field. Abrami (2001) states that besides SET's there is no other data readily available on teacher quality that is quantifiable and comparable. It might just be that SET's are what higher education institutions have settled for and are invested in continuing. Benton (2011) notes that RPT reviews might largely be based on faculty research and

scholarship and not instructor effectiveness because the preponderance of schools are not willing to do the relatively difficult, time-consuming work of actually studying the performance of faculty in the classroom. Schools presumably know how to do this – with tests of knowledge before and after a course to measure real growth in knowledge - but the reality of how much work it would be to do this regularly and across the University is daunting. SET's, which are primarily online, automated, and performed by students, are too easy to ignore. And so, it is very likely that SET's will be the only regular source of data on instructor effectiveness for the near future, which underscores the need for them to be regularly studied.

The evidence available indicates that the reliability and validity of SET's have been shown. Of course, given the bias of journals to publish significant findings, it may be that more studies are published that have positive results for the instruments under study. Some schools study the validity of their own SET's to ensure that they really understand how they perform and how they should be interpreted. But, at the same time there is also evidence that bias or the real possibility of bias exists in these kinds of instruments.

It is probably most accurate to say that there is no consensus on the validity of SET instruments because there are so many different ones in use. At the same time, SET's, regardless of their validity, are regularly being used for important faculty retention, promotion, and tenure decisions. The use of biased instruments could be considered discriminatory. Huston (2006) makes the point that if a business used an instrument to help make promotion decisions and there was no documented evidence that it was a fair test, or even the slightest evidence that is was unfair, there would probably be immediate lawsuits. Yet, at the same time, universities are using SET instruments that often do not

have any research behind them to show that they are unbiased. This is reason enough to urge institutions to undertake the regular study of SET's. Gravestock and Gregorleaf (2008) make a recommendation that fits well here: "Approved instruments should be evaluated by experts in survey construction and continuously investigated through institutional research" (p. 45), and if they can't, they suggest that the institution use one of the commercially available instruments (Gravestock & Gregorleaf, 2008).

SET's at the University of Utah

At the University of Utah, SET's are referred to as "Student Course Feedback (SCF)." At the end of every course, but before they see their grades, students are asked to evaluate the course and the instructor by indicating their agreement with seven course and seven instructor statements (see Appendix A). These questions are asked in a way such that students can respond by simply checking one of six boxes which correspond to the following ratings: Strongly Disagree, Disagree, Somewhat Disagree, Somewhat Agree, Agree, and Strongly Agree. If a student responds to the SCF questions and the instructor has turned in grades for the course then the student can see grades earlier than the date on which grades are published for all courses. Each semester, the University of Utah has approximately 70% of their students complete the SCF, a response rate that is considered one of the highest in the nation.

The SCF system is administered by a program manager housed in the University's Center for Teaching and Learning Excellence (CTLE), which is, in turn, overseen by the Office of Undergraduate Studies. The University has a contract with an outside vendor

(Smart Evals) to deliver SCF's online and to provide instructors, administrators, and staff with interfaces to access results.

History of SCF at Utah - The University has done some form of course evaluation for the past twenty years or so. Initially, departments conducted their own evaluations and they were very different across campus. These evaluations were not made public. In the mid-90's, the student association asked the faculty to see the results of those evaluations. The faculty said that they would do so but that they would like to first create standardized questions for all students (departments were still allowed to include their own questions in addition to the new ones). It was at this time that the 14 standardized questions now in use were created (see Appendix A).

There is no existing history of how these questions were chosen or what teaching constructs the instrument was attempting to measure. One can assume that the authors of the instrument did have some constructs in mind when drafting the original questions. If only a general impression of the course and instructor for summative purposes was desired, they could have just asked general outcome-related questions such as C2, C6, C7, and I7 (see Appendix A). That the authors asked 14 questions seems to indicate that they were interested in producing specific feedback that could be used by faculty and administration for formative purposes. The specificity of the non-summative questions from the list of questions supports this idea (Appendix A). For example, question I4 asks to what degree "The instructor created/supported a classroom environment that was respectful" and C5 asks about whether "Assignments and exams reflected what was covered in the course." As mentioned earlier, Cashin (1995) says that SET instruments

commonly ask questions around constructs that are salient in the classroom: organization, instructor skills, instructor-student rapport, etc.. The University of Utah's SCF instrument appears to align with this custom.

No known attempt has been made to validate the design of Utah's SCF instrument for whatever the specific purposes that were intended. In the process of developing any instrument, an analysis of the performance of the items is appropriate and recommended regularly to ensure the instrument's relevance and validity. It is very possible, for instance, that many of the questions are highly correlated with each other and represent one latent factor. If the designers were interested in developing a test with as much parsimony as possible, it might make sense to reduce the number of questions that load highly onto a particular factor if there is no other use for the redundant questions. It is, of course, possible that individual items are very useful to certain professors or departments and used by them for formative purposes and they would not want individual items to be removed. On the other hand, some users of the instrument might be interested in getting results from the instrument about particular facets of teaching that might require several items on the instrument to assess. This would be a good reason to keep several items per facet (or "construct").

Unless the instrument's factor structure and its measurement properties are analyzed, none of the uses described above can be validated. In addition, unless the instrument's construct validity can be demonstrated to be invariant, comparisons across certain population groups within that factor structure will also not be valid. The current study is an attempt to begin that process so informed decisions can be made about the potential desired uses described above.

SCF Factor Analysis Pilot Study - A pilot study using exploratory factory analysis by the current author (St. Andre, 2010) of the responses of two classes to the SCF questions revealed that there was a large factor present behind the 12 questions (The pilot did not use the two "overall questions" C7 and I7. Also, a separate exploratory factor analysis was conducted for each class). One factor explained just over 75% of the variance in each of the separate analyses of the two classes. Items loaded onto the large factors in somewhat different ways for the two classes[1], but no confirmatory test was done to investigate whether any specific number of factors was present based on any proposed design of the instrument to capture several facets of teaching. No work was done to examine whether the factors were invariant across these two groups either – whether the items in those two factors loaded onto the large factor in similar ways, indicating that the factor "meant" the same thing across meaningful groups within the University.

The selection of the two classes in the pilot was largely based on convenience, so the findings are only exploratory, and not generalizable. However, the finding of one large factor for 14 questions begs the question as to what the factor structure might look like using a larger and more representative selection of respondents. Also, because the intended construct structure of the instrument was unknown, it was proposed by the current author in the pilot (St. Andre, 2010) that future studies of the instrument begin with a content analysis. The purpose of this analysis would be to separate the items into a

[1] The classes used for this analysis were from Fine Arts and Sciences – the one from Fine Arts was a large class, the one from Sciences was smaller, so data were pulled from multiple semesters to get a large enough sample size to do the exploratory factor analysis.

reasonable construct structure, which could then be tested using a confirmatory factor analysis procedure.

Construct Analysis - In order to examine the factor structure of the instrument, the content of the SCF questions were analyzed in order to determine how they might be divided up into constructs. Based on the organization of the instrument into seven course and seven instructor questions, the most obvious structure is one with two factors - one for course and one for instructor, with the respective questions loading onto each of them and the factors correlating

Table 1: Hypothesized Constructs in University of Utah's Student Course Feedback Instrument

Construct	Question (C=Course question, I=Instructor question)
1. Organization of Course and Materials	C1: The course objectives were clearly stated. C3: The course content was well organized. C4: The course materials were helpful in meeting course objectives. C5: Assignments and exams reflected what was covered in the course.
2. Course Outcomes	C2: The course objectives were met. C6: I learned a great deal in this course. C7: Overall, this was an effective course.
3. Effective Learning Environment	I4: The instructor created/supported a classroom environment that was respectful. I5: As appropriate, the instructor encouraged questions and opinions. I6: The instructor was available for consultation with students.
4. Instructor Skills	I1: The instructor was organized. I2: The instructor demonstrated thorough knowledge of the subject. I3: The instructor presented course content effectively. I7: Overall, this was an effective instructor.

A more detailed examination of the questions also reveals that within each of those constructs – course and instructor – there might also be two more levels that make sense: "Organization of Course and Materials", and "Course Outcomes" within the course

questions, and "Effective Learning Environment" and "Instructor Skills" within the instructor questions (see Table 1). Constructs 1, 3, and 4 are largely formative in nature while construct 2 is summative, asking the student about overall outcomes from taking the course.

The content analysis of the existing instrument, then, suggests the possibility of a two-factor (course and instructor) or a four-factor structure (see Table 1). The pilot study showed, at the very least, the existence of a substantial general factor in the items. The current study, therefore, will test which of a one-, two-, or four-factor structure best fits the data from the University of Utah's SCF instrument. These three models are nested: one can create the two-factor model from the one-factor model and the four-factor model from the two-factor just by adding lambdas (factor loadings) to the desired factors.

Research Questions - The current paper represents a first step in the examination of the validity of the Student Course Feedback (SCF) instrument that has been in use at the University of Utah since the mid-1990's. While a thorough validity analysis is beyond the scope of the current project, it will attempt to answer two related research questions:

> **Question 1**. Do the responses from students to the 14 course evaluation questions indicate that the instrument is measuring one factor, two factors, or four factors?
>
> **Hypothesis 1 –** The data will show that a four-factor model will have adequate fit and fit the data significantly better than the other two models. (see Model 3 in Figure 1).
>
> **Question 2**. Is the latent factor structure[2] of the instrument (using whichever model emerges from Question 1 as the most reasonable) comparable across college, gender of instructor, and ethnicity of instructor? In other words, is there

[2] If none of the three models tested in Question 1 have adequate fit then Question 2 will not be addressed, as there is no sense in examining the measurement invariance of a poor-fitting model.

dimensional, configural, and item invariance in the SCF instrument, which would allow for a comparison of factor and/or item means across the groups of interest?

Hypothesis 2 – The instrument will have dimensional invariance - the construct structure will be the same across the different categories of the groups. For example, a confirmatory factor analysis model performed on courses taught by male instructors will show the same fit as one conducted on data for female instructors. At the item level, however, it is hypothesized that the form will vary by population – that items will load onto the factors in significantly different ways.

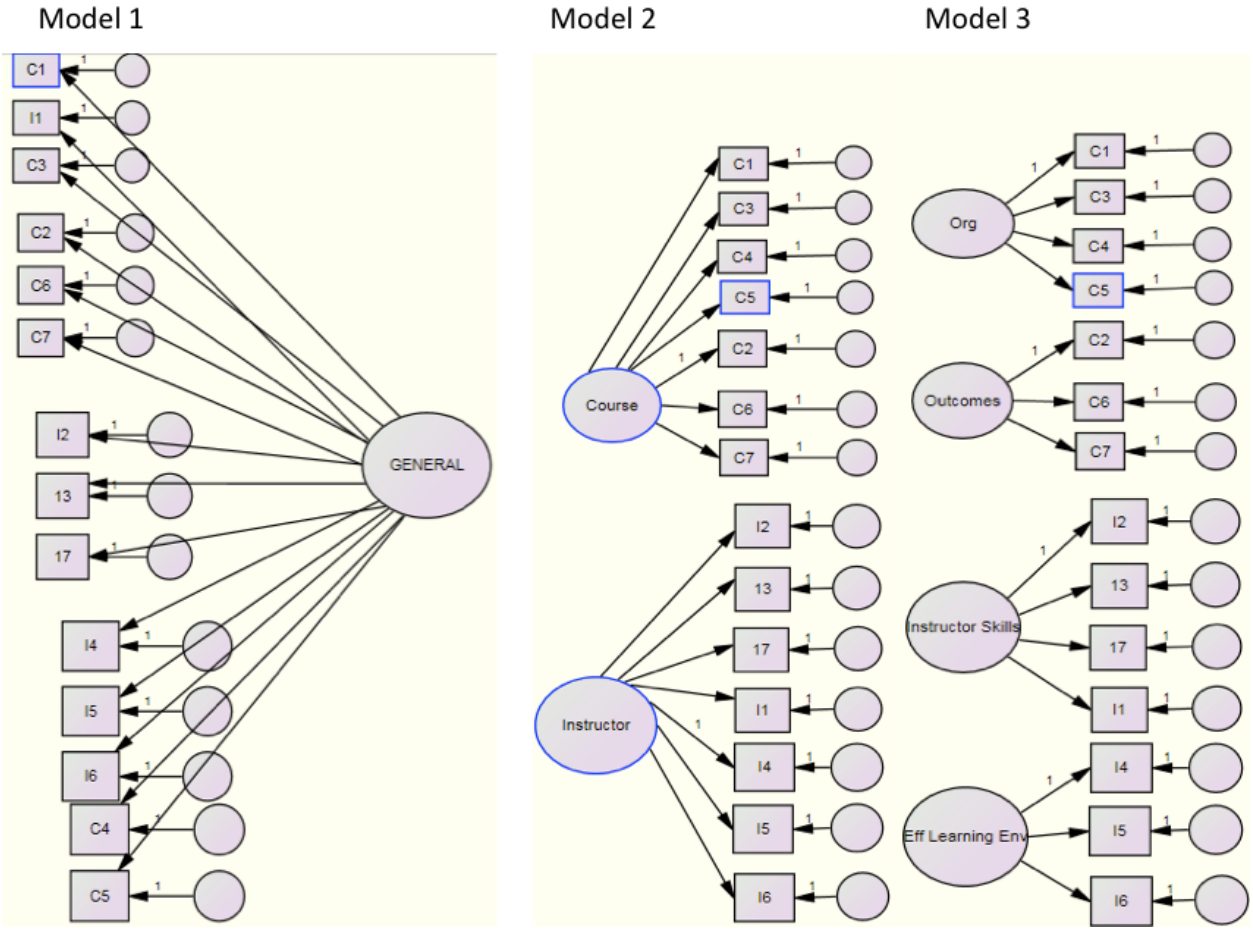

Figure 1: Three Competing Factor Models for the University of Utah's SCF Instrument (see Appendix A for a description of each indicator/question)

METHODOLOGY

This section will describe the data that was used and the method of analysis for each of the research questions.

Data - The data used in this study were the SCF responses of virtually all of the students who took courses in 13 departments selected from 6 colleges during the Spring 2009 semester at the University of Utah (see Table 2). Only those students who completed all of the questions were included in the study. This amounted to 23,516 responses. There are, of course, many students who responded to the questions more than once because they took multiple courses during the semester in these 13 departments. The departments were selected because of their size and their significant contributions to the undergraduate mission of the university.

Table 2: Colleges and Departments In Analysis

College	Department
Business	Management
	Marketing
Engineering	Mechanical Engineering
Fine Arts	Dance
	Music
Humanities	Communication
	English
	Writing
Social and Behavioral Sciences	Political Science
	Psychology
Science	Biology
	Computer Science
	Mathematics

Subjects - The subjects in the study were the students generating the responses to the questions. Instructor gender and ethnicity for each course were obtained from the

University's Office of Budget and Institutional Analysis (OBIA) and merged with the course evaluation data to create the dataset.[3] There were also, in some cases, multiple instructors per course and they were all included in the analysis.

This project has been approved by the Institutional Review Board at the University of Utah.

Question 1

This question asks which of the three models best fit the data.

Model 1 – A general factor that loads onto all 14 questions.
Model 2 – A two-factor model: one for the course questions and one for the instructor questions.
Model 3 – A four-factor model that loads as shown in Figure 1.

Model Fit - To determine which of these models best fit the data, three confirmatory factor analyses were conducted, one for each model. Because of the non-normal distribution of responses to the SCF items, a weighted least squares estimator was used in the CFA rather than maximum likelihood. Because parsimony is desired, if two models have comparable fit then the simplest model will be used for the subsequent analysis of measurement invariance. Also for reasons of parsimony, the examination of the models will begin with the simplest, Model 1, and proceed to the most complex, Model 3. The comparison of models will proceed as follows.

Because these models are nested, the optimal way to compare their fit is to conduct a chi-square difference test between them. Measurement models are nested when parameters are only added or removed to create the subsequent model. The chi-square

[3] The author would like to thank Joyce Garcia, Camille Wintch, Jill Stephenson, and James Anderson for their help and guidance in assembling the data needed for this study.

difference test compares the fit of the two models by taking the difference in the chi-square fit statistics and using the difference in their degrees of freedom and testing that against the chi-square distribution. Because of the large sample size in this study (24,000+ rows), and the sensitivity of the chi-square statistic to sample size, a new analytical procedure was needed because almost any difference would be found to be significant. As a result, Hoelter's Index was calculated on the differences between the chi-square statistics for each model.

Hoelter's test does not produce a significance value but instead produces a sample size below which the difference in the models would not be significant (Kenny, 2011). This provides a way of controlling for the impact of sample sizes when interpreting results. The formula for the Hoelter Index is:

$[(N-1)X^2 \text{ criterion}/(X^2)] + 1$

Because the meaningful group size for course evaluations at the University is the classroom, and the classroom size is rarely above 100 students, it was determined that the criterion for this analysis would be 100. Therefore, if the calculated Hoelter Index for a chi-square difference is larger than 100 that difference will be considered not meaningfully significant.

Model 1 will be compared to Model 2 and the best fitting model will be compared to Model 3. The resulting model with the best and adequate fit will then be used for the second question, which examines measurement invariance of the instrument. If none of these models has adequate fit then it does not make sense to test measurement invariance for that model.

Question 2

This question asks whether the instrument functions in comparable ways for courses in different colleges, and for instructors of different gender and ethnicity. The use of confirmatory factor analysis to examine how the latent factor structure of an instrument compares across different subgroups of a population is well-established (see Gregorich, 2007; Brown, 2006; Dimitrov, 2006). Showing measurement invariance of the overall model as well as at the factor and item level allows for comparisons across groups at those various levels to be made. One of the strengths of Multiple Group CFA is that all types of invariance across groups can be examined (Brown, 2006).

Measurement Invariance - To test whether the instrument functions differently for those in different colleges, instructors of different genders, and ethnicities, a series of nested tests of measurement invariance will be conducted. However, in order to look at measurement invariance across groups with many categories such as colleges and ethnicities, a decision was made to limit the analysis to a couple of specific comparisons so that the number of tests did not grow too prohibitively large. Four of the six colleges were divided into two groups: those that are more quantitative (science and engineering) and those that are more qualitative (fine arts and humanities) and the analysis will be done on those two groups. A similar issue arises when considering the comparison of instructors across all ethnic groups. In this case, the comparisons will be conducted between white and foreign instructors. Anecdotal evidence suggests that foreign (international) instructors are evaluated in very different ways to white instructors, presumably because of the likelihood that instructors speak English as a second language to their native

language and this contribute to a different classroom experience. It is also interesting to note that "foreign" is the second largest self-identified ethnic group at the University[4]).

The examination of measurement invariance requires a series of nested factor analyses to be conducted:

1. **Dimensional Invariance** is typically the first step tested by examining whether the instrument produces the same number of factors across the categories of the group of interest. Because the current analysis is testing the hypothesis that a certain model that was interpreted from the questions fits the data, it was considered important to test this same model for all groups, regardless of what an exploratory factor analysis might reveal. Therefore, dimensional invariance was assumed.

2. **Configural Invariance** will be examined by determining if the fit of the chosen model is different for those in the different categories in each group. E.g. the model fit for males will be compared to the fit for females to examine if they are both adequate. If the model does fit differently for different populations then the analysis will not proceed to the next step.

3. **Metric Invariance** will be examined by creating a stacked structural equation model using the group that is being compared as the "stacked" variable. Subsequent analyses are performed on nested models using a series of equality constraints.

[4] It seems inappropriate to label an ethnic group as "Foreign" because that term is non-specific and feels exclusive, but nonetheless is a standard term used in higher education institutional data.

In the first model, **Equal Forms**, the lambdas (factor loadings) between the factors and the items and the taus (intercepts of the items) are allowed to estimate freely and are not equated across the groups. This forms the baseline model.

In the second model, **Equated loadings,** the taus (intercepts) are left free to vary, but the lambdas are equated so it can be determined if the fit of the model suffers when the lambdas are not allowed to vary across the groups. If it does significantly suffer then it can be said that the measures "want" to vary between the categories of the group, and by not allowing it the fit of the model suffers. As Brown (2006) says "The test of equal factor loadings is a critical test in multiple groups CFA...this test determines whether the measures have the same meaning and structure for different groups of respondents" (p. 279; Brown, 2006).

The third model, **Equated Indicator Intercepts** is evaluated by holding both the lambdas and the taus to equality across the groups to determine if the intercepts (means) of the indicators also "want" to vary across the two groups. Lack of invariance here would mean that the means are different for the indicators of each latent variable.

RESULTS

Before addressing the research questions, a breakdown of the demographics for quantitative and qualitative colleges is given in Table 3a and 3b to show the distribution of the instructor demographics by college type. There are a greater percentage of men in the quantitative (science and engineering) than qualitative (fine arts, humanities) colleges. Foreign instructors[5] are more prevalent in the quantitative colleges as well. Women are more highly represented in the qualitative colleges.

Table 3a: Gender and Ethnicity (White/Foreign) of Instructors in Qualitative (Fine Arts, Humanities) Colleges

		Gender		
		Male	**Female**	**Total**
Ethnicity	**White**	3830 (58.3%)	2535 (38.6%)	6365 (97.0%)
	Foreign	156 (2.4%)	44 (.7%)	200 (3.0%)
	Total	3986 (60.7%)	2579 (39.3%)	6565 (100%)

Table 3b: Gender and Ethnicity (White/Foreign) of Instructors in Quantitative (Science, Engineering) Colleges

		Gender		
		Male	**Female**	**Total**
Ethnicity	**White**	6086 (68.7%)	1327 (15.0%)	7413 (83.7%)
	Foreign	1201 (13.6%)	244 (2.8%)	1445 (16.3%)
	Total	7287 (82.3%)	1571 (17.7%)	8858 (100%)

Table 4 shows overall means for the instructor items and course items. Instructors and courses in qualitative colleges are rated more highly than those in quantitative

[5] "Foreign" is a category in the ethnicity codes of the University's information systems and is used by faculty to indicate they were not born in the United States. There is some controversy in this practice, but it does allow us to test the hypothesis that foreign instructors, who are more likely to have English as a second language, are rated in different ways than others. The "White" category was chosen to compare them to as it was seen as the best option available for comparing foreign ESL instructors to native English speakers, although the author recognizes this is by no means an unassailable truth.

colleges. Females are rated more highly than males, and white instructors are rated more highly than foreign instructors. This may be an artifact of there being more females than males in those colleges.

Table 4: Average Scores (and Standard Deviations) for Course and Instructor Questions by College Type, Gender and Ethnicity of Instructor*

	Average of 7 Course Questions (SD)	Average of 7 Instructor Questions (SD)
College		
Qualitative (Humanities & Fine Arts) (n=7,330)	5.27 (.96)	5.39 (.87)
Quantitative (Science & Engineering) (n=9,568)	5.05 (1.07)	5.17 (1.01)
Gender of Instructor		
Male (n=16,376)	5.12 (1.04)	5.25 (.96)
Female (n=7,140)	5.19 (1.00)	5.32 (.93)
Ethnicity		
White (n=18,481)	5.16 (1.02)	5.29 (.94)
Foreign (n=1,973)	5.01 (1.08)	5.05 (1.01)

*aAll differences between categories of groups (quantitative vs. qualitative colleges, male vs. female, white vs. foreign) were significant at p<.001, effect sizes all less than .05.

The results of the analysis of the two research questions are presented below.

Question 1 – This question asks whether Model 1, 2, or 3 best fits the data, and the comparison of the models starts with the most parsimonious – the one factor Model 1 – compared to Model 2, the course and instructor items loading onto a course and instructor factor respectively. To make this comparison, a confirmatory factor analysis was performed on each of these models and the fit indices for those analyses are presented in Table 5a.

The chi-square goodness of fit indices were significant at p<.001 for both models, which typically indicates that the fit of the model is not good (see Table 5a). However, this

index is adversely affected by large sample sizes, and the sample size in this study is 25,316, so this figure is not considered appropriate in this application.

The Root Mean Square Error of Approximation (RMSEA) is a non-centrality based fit index, which examines how much discrepancy per degree of freedom there is between the observed covariance matrix and the one implied by the model (Butner, 2005). It penalizes for unnecessary parameters in the model. It runs high for low sample sizes, and is ideally less than .05. The scores of .204 and .140 in Models 1 and 2 both indicate poor fit but that could be the result of there being too many parameters (items, in this case) in the model.

The Comparative Fit Index (CFI) is also a noncentrality-based index, which has relatively small standard errors and is only minimally affected by sample size. Model 1 had a CFI of .826 and Model 2 was .919. Values of .95 or higher are considered good fit and above .90 are considered adequate.

The Standardized Root Mean Square Residual (SRMR) is an absolute fit index that is the average of the non-diagonal elements from a standardized residual correlation matrix and a value lower than .08 is considered adequate and below .05 is considered good. The fit of Model 2 (.028) appears to be good in this case and Model 1 (.054) to be only adequate.

The Bayesian Information Criterion (BIC) index is another absolute fit index. Its value is calculated as such: X^2 +[k(k-1)/2-d.f.] ln(N). The value of the BIC itself has no inherent meaning but it can be used to compare the fit of non-nested models. The smaller value indicates a better-fitting model. There is also a sample-size adjusted BIC that takes into account sample size. Because of its ability to compare non-nested models and take into account sample size it is utilized here. The value of Model 2 (594384) compared to Model 1(634600) again indicates that Model 2 is a better fitting model.

Finally, the difference between the models was evaluated using the Hoelter Index. As a reminder, the Hoelter Index produces a sample size, below which the difference between the models is considered not meaningfully significant. The meaningful group size in this study is the classroom. Because the largest class size at the University is rarely over 100, the value of 100 was set as the criterion for this analysis. Therefore, any significant chi-square differences with a calculated Hoelter Index above 100 will be considered to be differences that are the result of sample size and not necessarily a "meaningfully" significant difference.

Table 5a: One vs. Two-Factor Confirmatory Factor Analysis Model Fit Statistics

Model	**X^2 (d.f.)**	**RMSEA**	**CFI**	**SRMR**	**BIC (sample size adj)**
One Factor (1)	75180* (77)	.204	.826	.054	634600
Two Factors (2)	34957* (76)	.140	.919	.028	594384
X^2 Difference	40223** (1)				

* p<.001, n=23516

** p<.001, n=23516, Hoelter Index of this significant chi-square difference=3.25. Therefore, the difference between these models is meaningfully significant down to a sample size of 3. Since virtually every course is larger than 3 students, this difference is considered meaningfully significant.

The calculated Hoelter Index for the difference between the two-factor and the one-factor models in Table 5a was 3.25. This figure indicates that the difference between these chi-square values is meaningfully significant for the population in this study, where the largest class sizes are around 100 students. This statistic can be interpreted to mean that the two-factor model represents a significant improvement of fit over the one-factor model.

Because of this finding, the analysis proceeded to a comparison of the two-factor model to the four-factor model (see Table 5b).

The comparison of the two-factor to the four-factor model indicates that the four-factor model is an improvement of fit over the two-factor model. The RMSEA of the four-factor (.124) is lower, the CFI (.941) is higher and is a good fit, and the SRMR of .030 is very similar to the .028 in the two-factor model. The BIC is also lower (585024 vs. 594384). But, most importantly, the chi-square difference between the models is significant ($p<.001$) and the Hoelter Index calculated on the chi-square difference between the models is 28.7. Again, because this figure is less than the criterion of 100, this difference is considered meaningfully significant and the four-factor model can be interpreted to be an improvement in fit compared to the two-factor model.

Table 5b: Two vs. Four-Factor Confirmatory Factor Analysis Model Fit Statistics

Model	X^2 (d.f.)	RMSEA	CFI	SRMR	BIC (sample size adj)
Two Factors (2)	34957* (76)	.140	.919	.028	594384
Four Factors (4)	25563 * (71)	.124	.941	.030	585024
X^2 Difference	9394** (5)				

* p<.001, n=23,516

** p<.001, n=23,516. Hoelter Index=28.7. The difference between these models is meaningfully significant down to a sample size of 28. Since the majority of courses are of a size 29 or higher (median class is size is 29 and many smaller courses are independent studies or thesis hours, etc.), this difference is considered meaningfully significant.

The four-factor model appears to be the best fit for the data of the three models tested. The text of each question and standardized factor loadings are presented in Table 6 below. These loadings are also represented in Figure 2, which shows a full drawing of the four-factor model, the loadings, covariances between factors, and the residual variances of the items.

Table 6: Four-Factor Model Structure, Questions, and Factor Loadings for University of Utah's Student Course Feedback Instrument

Construct	Question (C=Course question, I=Instructor question)	Factor Loadings in Four Factor Model*
1. Course Organization	C1: The course objectives were clearly stated.	.875
	C3: The course content was well organized.	.904
	C4: The course materials were helpful in meeting course objectives.	.896
	C5: Assignments and exams reflected what was covered in the course.	.869
2. Course Outcomes	C2: The course objectives were met.	.892
	C6: I learned a great deal in this course.	.939
	C7: Overall, this was an effective course.	.969
3. Instructor Skills	I1: The instructor was organized.	.875
	I2: The instructor demonstrated thorough knowledge of the subject.	.832
	I3: The instructor presented course content effectively.	.925
	I7: Overall, this was an effective instructor.	.959
4. Effective Learning Environment	I4: The instructor created/supported a classroom environment that was respectful.	.910
	I5: As appropriate, the instructor encouraged questions and opinions.	.922
	I6: The instructor was available for consultation with students.	.859

* All factor loadings were significant at p<.001

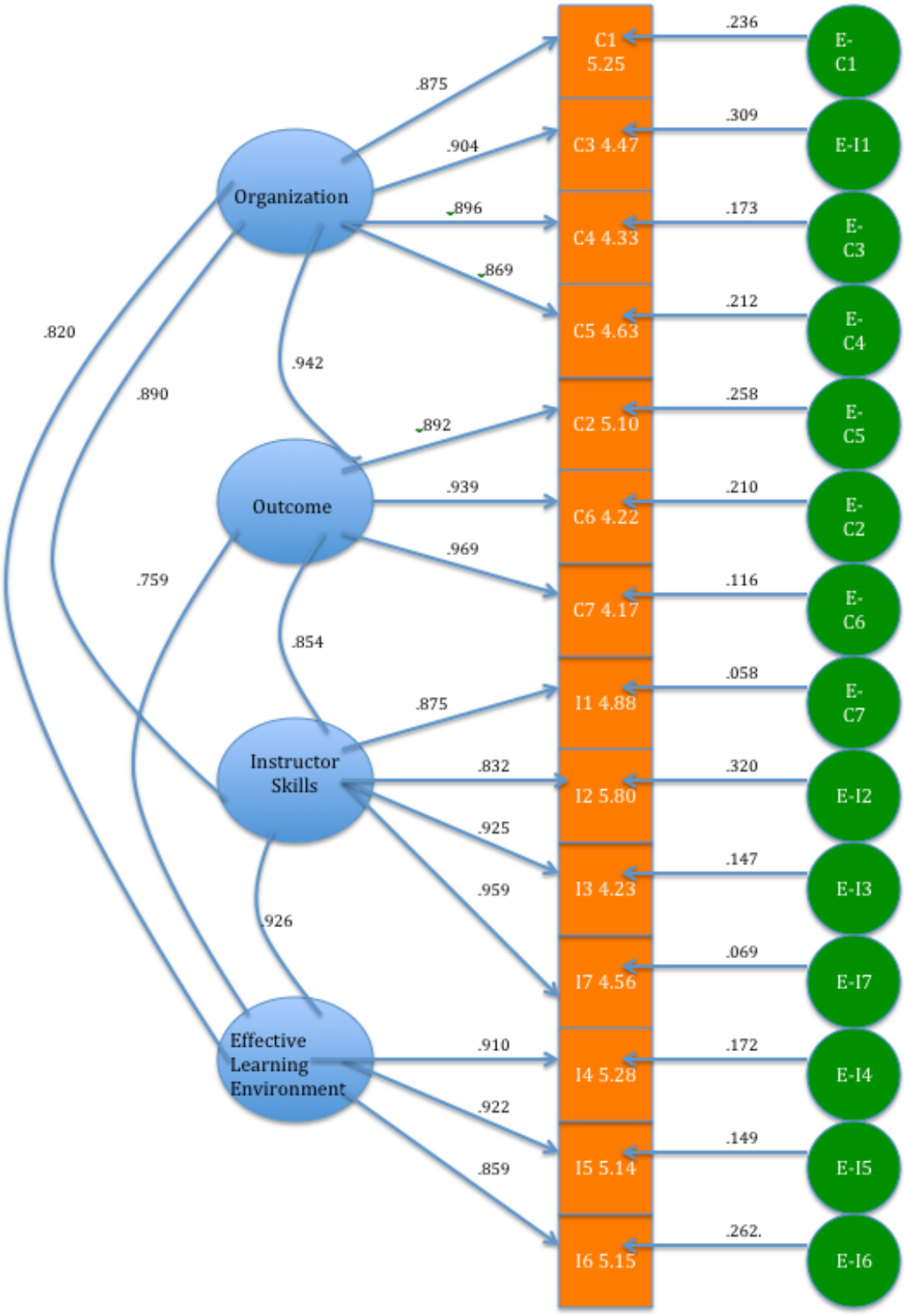

Figure 2: Four Factor Confirmatory Factor Analysis Model with Intercepts and Standardized Factor Loadings, Covariances, and Residual Variances

Question 2

The analysis of Question 2 will proceed with Model 3 – the four-factor model - because it demonstrated good fit and a significant improvement of fit over both the other models. The analysis of this research question will proceed through the various steps of measurement invariance across the three groups of interest: college type (qualitative vs. quantitative), gender, and ethnicity (white vs. foreign).

Table 7: Tests of Measurement Invariance of Student Course Evaluation Instrument by College: Qualitative (Humanities and Fine Arts) vs. Quantitative (Engineering and Sciences) (total n=16,898)

Test	X^2 (d.f.)	Hoelter Index	RMSEA	CFI	SRMR	BIC (sample size adj)
Model Fit by Category						
Qualitative	9762* (71)		.136	.928	.039	167239
Quantitative	1300* (71)		.043	.921	.047	253843
Form and Metric Invariance						
Equal Form	18862* (142)		.125	.910	.030	416187
Equal Factor Loadings	19087* (152)	161**	.121	.939	.037	416347
Equal Factor Loadings and Intercepts	19388* (162)	168**	.119	.938	.040	416581

*p<.001

** Not Significant. The Hoelter Index is > 100, which is the criterion for this analysis. This indicates that the difference between the model with these constraints and the one above it in the table is not meaningfully significant.

1. **Dimensional Invariance**. As mentioned earlier, for the purposes of comparing the four-factor model across groups it was assumed that the number of factors was the same across groups.

Table 8: Tests of Measurement Invariance of Student Course Evaluation Instrument by Gender (n=23,516)

Test	X^2 (d.f.)	Hoelter Index	RMSEA	CFI	SRMR	BIC (sample size adj)
Model Fit by Category						
Male (n=16,376)	22658 (71)**		.139	.924	.043	420466
Female (n=7,140)	9533** (71)		.137	.930	.039	169614
Form and Metric Invariance						
Equal Form	25939* (142)		.124	.940	.030	583902
Equal Factor Loadings	26204* (152)	164**	.121	.940	.038	584099
Equal Factor Ladings and Intercepts	26414* (162)	172**	.117	.939	.041	584240

*p<.001, n=23,516

** Not Significant. The Hoelter Index is > 100, which is the criterion for this analysis. This indicates that the difference between the model with these constraints and the one above it in the table is not significant.

2. **Configural Invariance**. This step examines whether the fit of the four-factor model is comparable across the groups under study. The Model Fit by Category sections of Tables 7-9 show that the fit of the four-factor model was good to adequate across type of college (Table 9), gender (Table 10), and ethnicity (Table 11). All of the SRMR values were below .05 for the categories in each group and the CFI indices were all above .90. These results were interpreted as indicating that there was configural invariance in the model.

3. **Metric Invariance-Equal Forms**. The metric invariance was tested by first establishing a baseline model, referred to as "Equal Forms" in Tables 5-7, which is the fit of the four-factor model stacked across the categories of each group. The loadings and the intercepts of the items are free to vary in this baseline model. These CFA's showed that the four-factor model, when run simultaneously ("stacked") for each of the categories of the groups, also had good fit. All of the SRMR's were below .05 with the exception of the "Foreign" category in Ethnicity, where the SRMR was .071, which is still considered adequate because it is below .08. All of the CFI's were above .90. The M-Plus code to run this model is provided in Appendix B.

Table 9: Tests of Measurement Invariance of Student Course Evaluation Instrument by Self-Identified Ethnicity: White vs. Foreign

Test	X^2 (d.f.)	X^2 diff (Δd.f.)	RMSEA	CFI	SRMR	BIC (sample size adj)
Model Fit by Category						
White (n=18481)	25232** (71)		.138	.926	.039	457910
Foreign (n=1973)	3370** (71)		.153	.907	.074	53734
Form and Metric Invariance						
Equal Form	22917* (142)		.125	.939	.030	506077
Equal Factor Loadings	23055* (152)	163**	.121	.939	.034	506147
Equal Factor Loadings and Intercepts	23211* (162)	171**	.118	.939	.036	506236

*p<.001; n=20,554

** Not Significant. The Hoelter Index indicates that the difference between the model with these constraints and the one above it is significant if the sample size (in this case, number of students in the classroom) is above the Hoelter Index figure. Because the sample size criterion for this analysis was set at the largest expected class size of 100, this difference was found to not be meaningfully significant.

3b. **Metric Invariance – Equal Loadings**. The next step in testing measurement invariance is to examine metric invariance. This is done by equating the loadings across the models, but to continue to allow the intercepts to vary. This model is then tested against the baseline model using the chi-square difference test. This test indicates whether the model suffers by constraining the lambdas. If the model fit decreases significantly it indicates that the lambdas "want" to vary in order to maintain the same model fit.

It is customary to evaluate this chi-square difference test of nested models by simply taking the difference in the chi-square values at these different levels of constraint and evaluating that chi-square value using the difference in degrees of freedoms. However, because of the very large sample size in this study and the fact that chi-square is sensitive to large sample sizes, a different analysis was done.

As described earlier, the Hoelter Index produces a statistic that represents the sample size below which the chi-square difference is not significant. If the sample size of the study is above the Hoelter Index figure then the difference between the models is considered significant. The meaningful group size in this study is the classroom. Because the largest class size at the University is rarely over 100, the value of 100 was set as the criterion for this analysis. If the Hoelter Index is larger than 100 then the difference will not be considered significant.

In all three of the groups under study the difference between the Equal Forms model and the Equal Loadings model was not significant for the Hoelter Index criterion of 100 set for the current study(see Tables 7-9). All of the Hoelter Indexes calculated on

the differences were larger than 100. This indicates that across college type, instructor gender, and ethnicity (white vs. foreign), the loading of items onto the four factors does not vary. Table 10 below shows the loading of items onto the factors across college, gender, and ethnicity.

Table 10: Standardized Factor Loadings of Items Across Categories of College, Gender, and Ethnicity

NEEDS redoing

		College		Gender		Ethnicity	
Factors	**Items**	**Qualitative**	**Quantitative**	**Male**	**Female**	**White**	**Foreign**
Organization	C1	.877	.871	.871	.833	.874	.862
	C3	.910	.908	.908	.915	.911	.899
	C4	.885	.890	.887	.889	.886	.892
	C5	.872	.847	.857	.873	.862	.852
Outcomes	C2	.882	.897	.893	.878	.889	.878
	C6	.944	.935	.941	.940	.940	.947
	C7	.973	.967	.970	.971	.970	.972
Instructor Skills	I2	.820	.827	.809	.860	.818	.841
	I1	.842	.815	.822	.853	.842	.723
	I3	.920	.927	.923	.924	.923	.924
	I7	.960	.970	.966	.962	.963	.971
Eff. Learning Environment	I4	.908	.911	.907	.916	.909	.910
	I5	.917	.928	.920	.930	.921	.943
	I6	.860	.857	.855	.871	.860	.855

3c. **Scalar Invariance – Equal Intercepts**. Because there was metric invariance, the analysis of measurement invariance can continue to examine scalar invariance – the equality of intercepts. This step reveals whether there are significant differences in the means of the items as they are represented in the four factors of the model. The current analysis found invariance in the intercepts (see Tables 7-9). There was no meaningful difference in the means of items within the four factors across type of college, gender, and ethnicity (see Table 10).

DISCUSSION

Any study of the construct validity of a measurement instrument should continue at regular intervals for as long as the instrument is going to be used. Validity is an argument about the appropriateness of an instrument at a certain time and place and in certain conditions. One author who is distinguished in the area of test validity, Samuel Messick, would say that the study of validity is not an act, it is a process, and one that should be done on a regular basis. His holistic view of validity can be summarized by the following:

> The key validity issues are the interpretability, relevance, and utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of social consequences of their use (Messick, 1988; p. 33).

Along the lines of those important validity issues, the purpose of the current study was to examine the factor structure and measurement invariance of the University of Utah's Student Course Evaluation instrument. In order to interpret and understand the utility of the SCF instrument as a whole or the individual scores on items, the current study investigated the underlying factors in this instrument as well as examined its measurement invariance. These analyses begin to answer the question of whether the whole instrument is needed, whether scale scores can be created based on multiple factors being present, and whether the individual scores have the same meaning across relevant groups of individuals. All of these steps are important in determining the interpretability and relevance of SCF scores.

There is no record of any formal evaluation of the instrument's psychometric properties, so this study began with an interpretation of possible constructs being measured by the instrument. The study then proceeded with confirmatory factor analyses

of three competing factor models and an examination of the measurement consistency (invariance) of the selected model across selected populations at the University.

A pilot study (St. Andre, 2010) that used an exploratory factor analysis found one factor, although that study did not do any confirmatory factor analyses and was also performed on only two courses. It is reasonable to expect that a two-factor model might better explain the data than a one-factor model because the 14 questions are divided into 7 course-related and 7 instructor-related questions. This turned out to be true: the two-factor model represented an improvement over the one-factor model.

The current study went on to hypothesize that each of those two factors could be further split into two factors. A confirmatory factor analysis of these four correlated factors–Organization, Creating an Effective Learning Environment, Instructor Skills, and Course Outcomes (see Table 1)– had good-to-adequate fit, and better fit than both a one-factor and a two-factor model.

An important validity question to ask after establishing the structure of a model is whether the model applies equivalently to different populations. This is typically done by examining subgroups of the population *responding* to the instrument. However, in the current analysis, for privacy reasons, no demographic data were available for the students taking the instrument. However, in the current analysis of a course evaluation instrument, it is also important to know that the instrument is invariant across characteristics of the instructors who are being rated. The question being asked in a measurement invariance analysis is whether the items from the instrument measure the same constructs (factor structure) for each of the categories of these groups, and whether there are similar relationships of items to those constructs (factor loadings) across those categories (Brown, 2006).

In the current study, the investigation of measurement invariance specifically answers this question: Does the instrument have the same number of factors (dimensional invariance), configuration of items into factors (configural invariance), loadings of items onto factors (metric invariance), and item means within those factors (scalar invariance) across instructors from different kinds of colleges, gender, and ethnicity.

Six separate CFA's showed that the four-factor model (which was assumed for the measurement invariance analysis) fit the two categories in each of the three groups equally well. This was interpreted to mean that regardless of the type of college the course was offered in, instructor gender, or instructor ethnicity (white vs. foreign), the items loaded onto the four factors in patterns that were not significantly different. The fit of the four-factor model was good-to-adequate for the six subgroups. This represents configural variance.

Once configural invariance was established, the next measurement invariance question was whether the loadings of the items onto each of those factors were comparable across the categories of each group (quantitative vs. qualitative colleges, men vs. women, white vs. foreign instructors). Across the board, this was found to be true. When the factor loadings (coefficients between items and factors) were forced to be the same, the fit of the model did not significantly decrease across the categories of any of the characteristics under study.

This finding can be interpreted to mean that scores produced by the Student Course Feedback instrument for each of the categories of these groups can be interpreted on the same scale, which is referred to as metric invariance. For example, an average score of 5.1 out of 6 on a particular question for a male instructor can be said to mean the same thing as

a 5.1 from a female instructor. This is an important finding, given that instructors are required to take one score from the SCF (I7: "Overall this was an effective instructor") for each course they teach and place it in their faculty HR file, where it can be assumed it will be used to compare instructors to each other.

It is unlikely that the original authors of the SCF intended it to exactly produce the four factors that were identified in the current study, or that the factor structure should necessarily be compared across the groups identified here. For the purposes of the current study the constructs were hypothesized and then analyzed as an exercise in demonstrating the kinds of analyses that are possible with CFA and to begin a discussion about the SCF instrument's construct structure. These findings and process can provide the University's Student Course Feedback Oversight Committee with ideas as to how they might interpret or analyze the instrument's factor structure in the future. Some suggestions include:

- The intended constructs of the instrument should be determined and made public so that their existence can be validated through confirmatory factor analysis and this process can be done as part of a regular assessment cycle.
- If the administration is interested in knowing faculty performance on particular aspects of teaching, then appropriate questions should be used (or newly written) to approximate those aspects, and then CFA should be used to verify their existence in student responses.
- If some questions do not load onto factors in a useful way and members of the faculty are not regularly using those questions, then they should be removed.
- A process should be put in place to evaluate the instrument for its purposes and changed if it's not meeting those purposes.

The SCF instrument, as are most student evaluations of teaching, is a hotly debated, if not polarizing, topic on the University of Utah campus. Many people write it off as only measuring characteristics of instructors that are unrelated to teaching, while others think it is simply correlated with grades, with the implication that some faculty partake in a trade

of good evaluations for good grades. At the same time, an instructor's average on the last question on the instrument, "Overall this was an effective instructor," continues to be required for the faculty personnel file, without any real understanding about what worth or validity that one statistic has. Also, no concerted effort has been undertaken to examine any potential bias based on gender or ethnicity might be present in the instrument. The current study takes a step in that direction by finding that SCF does not perform differently (does not represent a different scale of measurement) for courses taught in different colleges or for instructors of different genders or ethnicities (white and foreign). Hopefully, this paper is the beginning of a process that the University of Utah undertakes to regularly examine some of these important validity issues in the instrument.

APPENDIX A - UNIVERSITY OF UTAH STUDENT COURSE FEEDBACK QUESTIONS

Standard Course Questions

1. The course objectives were clearly stated.
2. The course objectives were met
3. The course content was well organized
4. The course materials were helpful in meeting course objectives
5. Assignments and exams reflected what was covered in the course.
6. I learned a great deal in this course.
7. Overall, this was an effective course.

Standard Instructor Questions

1. The instructor was organized.
2. The instructor demonstrated thorough knowledge of the subject.
3. The instructor presented course content effectively.
4. The instructor created/supported a classroom environment that was respectful.
5. As appropriate, the instructor encouraged questions and opinions.
6. The instructor was available for consultation with students.
7. Overall, this was an effective instructor.

APPENDIX B

Relevant Portion of M-Plus Code for Tests of Measurement Invariance for College Type

To test equal forms – lambdas (loadings) and taus (intercepts) freely estimated in both groups

```
…
    GROUPING IS Coll_Bin (0=qualitative 1=quantitative);

    ANALYSIS:
      TYPE IS GENERAL
      ESTIMATOR IS WLSMV;
      ITERATIONS = 10000;
      CONVERGENCE = 0.00005;

    MODEL:
    org by C1 C3*1 C4*1 C5*1;
    outcome by C2 C6*1 C7*1;
    instskls by I2 I1*1 I3*1 I7*1;
    lrnenv by I4 I5*1 I6*1;

    MODEL quantitative:
    org by C3*1 C4*1 C5*1;
    outcome by C6*1 C7*1;
    instskls by I1*1 I3*1 I7*1;
    lrnenv by I5*1 I6*1;
    [C3 C4 C5 C6 C7 I1 I3 I7 I5 I6];

    OUTPUT: tech1;
```

To test equal loadings – lambdas equated across groups, taus (intercepts) estimated freely

```
…
    GROUPING IS Coll_Bin (0=qualitative 1=quantitative);


    ANALYSIS:
      TYPE IS GENERAL;
      ESTIMATOR IS WLSMV;
      ITERATIONS = 10000;
      CONVERGENCE = 0.00005;

    MODEL:
    org by C1 C3*1 C4*1 C5*1;
    outcome by C2 C6*1 C7*1;
    instskls by I2 I1*1 I3*1 I7*1;
    lrnenv by I4 I5*1 I6*1;

    MODEL quantitative:
    [C3 C4 C5 C6 C7 I1 I3 I7 I5 I6];

    OUTPUT: tech1;
```

To test equal intercepts – lambdas and taus equated across groups

```
…
```

```
GROUPING IS Coll_Bin (0=qualitative 1=quantitative);

ANALYSIS:
  TYPE IS GENERAL;
  ESTIMATOR IS WLSM;
  ITERATIONS = 10000;
  CONVERGENCE = 0.00005;

MODEL:
org by C1 C3*1 C4*1 C5*1;
outcome by C2 C6*1 C7*1;
instskls by I2 I1*1 I3*1 I7*1;
lrnenv by I4 I5*1 I6*1;

OUTPUT: tech1;
```

REFERENCES

Abrami, P.C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. In M.Theall, P.C. Abrami, and L.A. Mets (Eds.). The student ratings debate: Are they valid? How can we best use them? [Special issue]. *New Directions for Institutional Research 109*, 59-87.

Ambady, N. & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64,* 431-441.

Benton, R. (2011). Using student course evaluations to design faculty development workshops. Academy of Educational Leadership Journal, 15(2).

Butner, J. (2005). Structural Equation Modeling Class Power Point Slides. University of Utah. Salt Lake City.

Butner, J. (2012). An RMSEA sample size adjusted chi-square difference test: SPSS syntax and data file. University of Utah. Salt Lake City.

Cashin, W.E. (1995). Student ratings of teaching: The research revisited (IDEA Paper #32). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Centra, J.A. (1993). *Reflective faculty evaluation. Enhancing teaching and determining faculty effectiveness*. San Francisco, Jossey-Bass.

Centra, J.A., Gaubatz, N.B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education, 71(1)*, 17-33.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education, 31 (16)*.

Delgado, R. & Bell, D. (1989). Minority law professors' lives: The Bell-Delgado survey, *Harvard Law Review, 24,* 349-354.

Feldman, K.A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education, 30*, 137-74.

Feldman, K.A.(1997). *Identifying exemplary teachers and teaching: Evidence from student ratings*. In R.P. Perry and J.C. smart (eds.), *Effective teaching in higher education*: Research and Practice. New York: Agathon Press.

Fischer, J.D. (2004). The use and effects of student ratings in legal writing courses: A plea for holistic evaluation of teaching. *Journal of Legal Writing, 10,* 111-112.

Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. In K.G. Lewis (Ed.), Techniques and strategies for interpreting student evaluations [Special issue]. *New Directions for Teaching and Learning*, *87*, 85-100.

Gardner, M. K. (2011). Theories of intelligence (pp. 79-100). In Bray, M. A., & Kehle, T. J. (Eds.) Oxford Handbook of School Psychology. New York: Oxford University Press.

Gravestock, P. & Gregor-Greenleaf, E. (2008). *Student Course Evaluations: Research, Models, and Trends.* Higher Education Quality Council of Ontario, Toronto.

Greenwald, A.G. & Gilmore, G.M.(1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52,* 1209-1217.

Gregorich, S. (1996). Do self-report insruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factory analysis framework. Med Care, 44(11), S78-S94).

Gustafsson, J.E. & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research,* 28, 407-434.

Huston, T. A. (2006). Race and gender bias in higher education: Could faculty course evaluations impede further progress toward parity*? Seattle Journal of Social Justice, 4,* 591-601.

IDEA Center (2012). *Student Ratings of Instruction*, Kansas State University, http://www.theideacenter.org/category/our-services/student-ratings-instruction.

Kennedy, E.J., Lawton, L., & Plumlee, E.L. (2002). Bliss ignorance: The problem of unrecognized incompetence and academic performance. *Journal of Marketing Education, 24*, 243-252.

Kenny, D. (2011). Measuring model fit. http://davidakenny.net/cm/fit.htm. September 4, 2011.

Marsh, H.W. (1982). *Student Evaluations of Educational Quality*, University of Wisconsin, http://www.uww.edu/learn/seeq.php.

Marsh, H.W. (1982). The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. Applied Psychological Measurement, 6, 47-59.

Marsh, H.W. & Roche, L. (1992). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. Campbelltown, Australia: University of Western Sydney.

Martin, D.J. & Radmacher, S.A. (2001). Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis. *Journal of Psychology, 135(3)*, 259-268.

McKeachie, W.J. (1997). Student ratings: The validity of use. *American Psychologist, 52,* 1218-1225.

Merritt, D.J. (2008). Bias, the brain, and student evaluations of teaching. *St. John's Law Review, 82*, 235-287.

Messick S. (1988) The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Brainer & H.I. Braun (Eds.), *Test validity*. Lawrence Erlbaum Associates, Inc. Hillsdale, New Jersey.

Paswan, A.K. & Young, J.A. (2002). Student evaluation of instructor: A nomological investigation using structural equation modeling. *Journal of Marketing Education, 24*: 193-202.

Seldin, P. (1985). *Current practices in evaluating business school faculty*. Pleasantville, NY: Pace University.

St. Andre, M. (2010). *A factor analysis of the University of Utah's student course evaluation system.* A presentation to the University of Utah's Educational Psychology Learning Sciences Seminar Class. Salt Lake City, UT.