STUDENTS’ ASSESSMENTS OF INSTRUCTION: A VALIDITY STUDY

By

Alan Brian Socha

A Thesis
Submitted to the
Faculty of the Graduate School
of
Western Carolina University
in Partial Fulfillment of
the Requirements for the Degree
of
Master of Arts

Committee:

____ Director

____

____

____ Dean of the Graduate School

Date: _____________________________________

Spring 2009
Western Carolina University
Cullowhee, North Carolina

STUDENTS' ASSESSMENTS OF INSTRUCTION: A VALIDITY STUDY

A thesis presented to the faculty of the Graduate School of
Western Carolina University in partial fulfillment of the
requirements for the degree of Master of Arts in Experimental Psychology.

By

Alan Brian Socha

Director: Dr. David M. McCord
Department Head of Psychology
Psychology Department

Committee Members: Dr. Bruce B. Henderson, Psychology
Dr. John Habel, Psychology

April 2009

ACKNOWLEDGEMENTS

I would like to acknowledge several people who helped me complete this thesis. First, I would like to thank my thesis chair, Dr. David M. McCord, for his assistance, encouragement, and direction throughout this process. His guidance was invaluable and I would not have completed this process without his support. I also want to sincerely thank the other members of the thesis defense committee for their contributions and time, Dr. John Habel and Dr. Bruce B. Henderson.

Many colleagues and friends have also provided significant support, encouragement, and feedback throughout this process. I would like to especially thank Keith Stiles and Dr. Raymond Barclay for offering encouragement and/or advice when needed. Additionally, a special note of thanks goes to Western Carolina University for permitting me to conduct this research.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

ABSTRACT

STUDENTS' ASSESSMENTS OF INSTRUCTION: A VALIDITY STUDY

Alan Brian Socha

Western Carolina University (April 2009)

Director: Dr. David M. McCord

The aim of this study was to analyze the reliability and validity of the Activities, Independent Research, Internship/Practica/Clinical, Laboratory, Online, Seminar, Standard Lecture, and Studio-Performance course forms of the Students' Assessment of Instruction (SAI) instrument. The participants were volunteers from the student population at Western Carolina University (WCU) for the spring 2007, fall 2007, and spring 2008 semesters. The reliability and validity were analyzed using confirmatory factor analysis (CFA), generalizability theory (g-theory), and hierarchical linear modeling (HLM). Findings were mixed. The Independent Research Course Form had no evidence of reliability and validity while the Online Course Form had some evidence of validity but no evidence of reliability. All other course forms had moderate to excellent validity. Suggested changes for the course forms are discussed.

INTRODUCTION

Teachers can be effective or ineffective in many ways, thus there can be no single measure of effective instruction. A conceptually sound and properly implemented faculty evaluation system is therefore an important component for an effective school. A college or university's faculty evaluation system should be fair and contribute to the personal goals of the teacher (including personal and professional development), the mission of the instructional program, and the mission of the school.

A faculty evaluation system can be both summative and formative. It is summative when it is used for administrative decisions such as tenure, promotion, merit pay, and reappointment. It can be formative to the degree in that it can be used to improve instruction. Plans, syllabi, selection of supplemental materials, range and variety of alternative activities, and samples of testing forms and procedures are all examples of formative data.

Many teachers assert that the only individuals truly qualified to judge competency are colleagues (Wilson, Dienst, & Watson, 1973). Unlike students, colleagues are also able to judge course design and instructional materials. Such peer nomination and ranking methods are typically summative and offer little feedback to the teacher to improve his or her instruction (Cohen & McKeachie, 1980). Classroom observations can also be used to evaluate instructional effectiveness, and can be summative and formative. Observations are typically narrow in scope, e.g., they can be artificial, infrequent, and limited in focus not allowing an observer to evaluate the full repertoire of teacher duties and responsibilities (Stronge & Ostrander, 1997). Also, without evidence of student achievement no teacher can be in a position to assess the amount of student learning in a colleague's course. Students may be more qualified to evaluate many aspects of in-class behavior. This evidence is typically obtained through student responses on end-of-semester rating forms, which also can be summative and formative. Ultimately, a faculty evaluation system should utilize a range of methods since each has limitations (Macdonald, 1980b).

Of the many sources a college or university can use in their faculty evaluation system, students' assessments of instruction should be one of the primary sources because these assessments can help describe the learning environment more concisely. Students are the only individuals directly and extensively exposed to course elements and therefore are the most appropriate evaluators of the satisfaction and quality of those elements (Aleamoni, 1981). Students' assessments of instruction instruments should not ask students questions they cannot answer though. Judgments about course materials, how well the teacher knows the subject matter, and so on are best left to other sources in the faculty evaluation system.

Students' assessments of instruction have a large amount of appeal because they can be used to provide feedback on instructional effectiveness to faculty to facilitate self-improvement, can be used in personnel decisions, help improve effective instruction by increasing the likelihood that quality teaching will be recognized and rewarded, can be used to provide students with information to use in course and teacher selection, and can be used for research on teaching to answer questions like how teachers behave, why teachers behave the way they do, and what the effects are of teachers' behaviors (Kyriakides, Demetriou, & Charalambous, 2006; Macdonald, 1980b; Marsh, 1984, 1991, 1992; Marsh, Hau, Chung, & Siu, 1997; Thorpe, 2002). Unfortunately, the summative uses of students' assessments of instruction may give the assessments a threatening appearance to faculty, especially if there is no evidence of validity. This threatening appearance mostly comes from concerns on how students' inexperience and biases can affect their perception of instructional effectiveness.

Concerns about the appropriateness of students' assessments of instruction can cause disruption. Many teachers feel that students have a lack of experience and a lack of maturity and therefore cannot make consistent judgments about the teacher and instruction. There needs to be evidence that students are capable of using a reasonably appropriate weighting scheme and have insight into how they weight the individual teaching factors (Harrison, Douglas, & Burdsal, 2004). Also, current instructional methods are more varied than the items appearing on commonly used ratings instruments can support (Theall & Franklin, 2000). In order to address this variability in instructional methods, different instruments would be needed for each method of instruction.

Research on students' assessments of instruction has suggested that they are reliable and stable (Marsh, 1984, 1992). Reliability addresses the internal consistency, or error, in the ratings. Low reliability implies that there is random error. If an instrument is not reliable then the data and resulting evaluations will be meaningless (Aleamoni, 1981). Also, some teachers may be uniquely suited to teach specific courses, so the generalizability of ratings of different offerings of the same course by different teachers and offerings of different courses by the same teacher are important aspects of reliability.

Reliability makes up one component of construct validity, which has also been suggested to be a characteristic of students' assessments of instruction (Marsh, 1984, 1992). Some other components of construct validity are convergent validity, discriminant validity, and criterion-related validity (Kraiger & Teachout, 1990). Construct validity reflects the usefulness of a students' assessment of instruction instrument for measuring the students' view of effective instruction. Construct validity can never be completely present or absent, but assessing it can provide an understanding of the conceptual framework of the evaluation instrument. Convergent validity is the extent to which alternative methods provide similar information about ratees. Discriminant validity is the extent to which measures of theoretically distinct constructs are empirically related. Criterion-related validity is a reflection of the conceptual framework about a construct containing the construct's meaning, and reflects how well the underlying factor structure measures what the instrument intends to measure.

The validity of students' assessments of instruction instruments are sometimes discounted on the grounds that students tend to be too easily swayed by superficialities (Wilson et al., 1973).

These superficialities, or potential biases, make up most of the concerns faculty have about students' evaluations and give rise to the belief that students' evaluations may be unreliable and invalid. Potential biases can reflect validity if they have a similar influence on multiple indicators of instructional effectiveness, so validity research needs to carefully address each potential bias. Some potential biases are the students' expected grades, students' characteristics (e.g., gender, age, year in college, academic ability, subject matter interest), teacher characteristics (e.g., gender, enthusiasm or expressiveness, experience, research productivity, rank), course characteristics (e.g., class size, course requirements, course level, subject matter, topic difficulty), and the "halo effect" (current experiences affecting future experiences).

A series of committees consisting of faculty members created the Students' Assessment of Instruction (SAI) instrument at Western Carolina University (WCU). Different versions of the SAI were created for a variety of instructional methods. This instrument was created to obtain summative information for personnel decisions such as tenure, promotion, reappointment, and merit pay through administering the appropriate forms of the SAI across the university. Colleague reviews of teaching and a teacher's self-report and evaluation were to be used in conjunction with the SAI for that purpose. The second purpose of the SAI was to obtain formative information to use as a basis for making decisions for improving instruction. The SAI was designed based on the five factors of *Organization and Clarity*, *Enthusiasm and Intellectual Stimulation*, *Rapport and Respect*, *Feedback and Accessibility*, and *Student Perceptions of Learning*.

The current study analyzes the factor structure of the SAI using confirmatory factor analysis (CFA) to determine if the data reflect the five factors that were the basis for the instrument. This would be evidence of construct validity. The current study then uses generalizability theory to analyze the reliability of the SAI. Then hierarchical linear modeling (HLM) will be used to analyze other aspects of construct validity, including potential biases (e.g., students' characteristics, course characteristics, and faculty characteristics, with a larger focus on student's expected grade).

LITERATURE REVIEW

Definitions of Instructional Effectiveness

There is no single measure of effective instruction (Marsh, 1984; Travers, 1981). Marsh (1991) stated that "there are many ways to be an effective teacher and many ways to be an ineffective teacher" (p. 421). Teaching is a multidimensional activity (Algozzine et al., 2004) and criteria need to be defined for effective instruction in order to evaluate teaching (Macdonald, 1980a). Student learning, changes in student behaviors, teacher self-evaluations, peer/administrative evaluations, frequency specific behaviors observed by trained observers, and experimental manipulation effects are all accepted criteria of effective instruction (Marsh, 1984). Achievement is also a criterion because it is related to learning attitudes, values, appreciations, moral principles, cognitive perspectives, and other outcomes (Macdonald, 1980a). Faculty are expected to take on expanded roles and more responsibilities such as curriculum development, action research, team leading, and staff development facilitation (Kyriakides et al., 2006). Travers (1981) suggested that "teachers most effective in producing learning are clear in the expression of their ideas, variable and flexible in their approaches to teaching, enthusiastic, task-oriented, and so forth" (p. 19). Inconsistent operational definitions of instructional effectiveness make assessing effective instruction across contexts extraordinarily challenging. One's definitions of the goals of teaching define what are effective (McKeachie, 1997).

Some definitions of instructional effectiveness focus on aspects of the instructional process (d'Apollonia & Abrami, 1997). The focus of instruction has moved away from the memory of facts and definitions and toward the importance of the way knowledge is structured, and upon skills and strategies for learning and problem solving (Feldman, 1989). Preparation of course materials, provision of feedback, and grading are all aspects of the instructional process. Good teachers go beyond the textbook and the confines of the classroom. Out-of-classroom contributions are important because much learning occurs outside the classroom. A systematic view of what instructional activities teachers choose for their students and why, is a requirement for describing their instructional plans (Franklin & Theall, 1992). This view is also associated with instructional design and instructional development. Some instructional designs are better than others because they incorporate instructional activities that promote learning more effectively. Student learning and satisfaction are therefore associated with effective instructional designs because the most effective teachers are those who use methods and activities that best promote student learning and have the best repertoire of good teaching behaviors.

Whether or not a teacher sees teaching as the facilitation of learning or as the transmission of knowledge can influence the way he or she teaches and his or her choice of teaching methods (Willcoxson, 1998). Trigwell, Prosser, and Taylor (see Zhang, 2001) proposed two approaches to teaching. The first is a transmission/teacher focused approach. Teachers using this approach tend to be content-oriented and emphasize the reproduction of correct information. The second approach is the conceptual change/student focused approach. Teachers using this approach are learning-oriented and concerned with students' conceptual change and growth. Zhang (2001) found evidence that teachers taking the transmission/teacher focused approach tend to be engaged in such teaching activities as lecturing about facts and requiring students to reproduce what they have learned in detail. Zhang also found evidence suggesting that teachers taking the conceptual change/student focused approach tend to provide students with intellectual autonomy and chances to make their own decisions, create a learning atmosphere in which students are allowed to evaluate different viewpoints, and encourage students to focus on bigger pictures of the issues encountered in students' learning tasks.

Other definitions focus on products that effective instruction promotes in students (d'Apollonia & Abrami, 1997). Subject-matter expertise, skill in problem solving, and positive attitudes toward learning are all examples of the products that effective instruction promotes in students. Skills for continued learning and critical thinking, motivation for lifelong learning, and changes in attitudes and values can also be used as definitions for effective instruction.

Scriven (1980) suggested that professional teaching should contribute to substantial positive gains in student learning, such as in content, skills, and attitude. Student achievement should be related to the goals of instruction (Macdonald, 1980b). Also, learning gains from teaching should exceed learning gains simply achieved by reading a book.

Schrodt, Turman, and Soliz (2006) suggested that a variety of communication skills can enhance student learning and motivation. Perceived understanding involves students' assessments of their success or failure when attempting to communicate with the teacher. By confirming behaviors, teachers communicate to students that they are recognized, valued, and appreciated. Perceived understanding will therefore mediate the influence of confirmation, making teacher confirmation an important communication skill. Teachers communicate a sincere interest in students when they respond to questions in an inviting manner. This interactive teaching style fosters learning, invites student participation, and engenders feelings of success for students as they attempt to communicate with their teachers. In the end, this will enhance perceptions of teacher credibility and evaluation (Schrodt et al., 2006).

For students, effective instruction is teaching that they perceive as assisting their chosen approach to learning (Willcoxson, 1998). Some prefer teaching that is well-organized, allowing them to passively listen while simultaneously preparing them well for tests (McKeachie, 1997). This approach to learning is similar to Biggs's first learning approach (see Zhang, 2001). Biggs's first approach is a "surface approach," where students have a motive for obtaining a degree and use a learning strategy that allows them to do the minimum to get by. Biggs proposed two more learning approaches beyond the surface approach. The second is the "deep approach." This is where students have an intrinsic motive for learning and use a learning strategy that allows them to achieve a true understanding of the material learned. The third is the "achieving approach" where students have a motive for academically surpassing their peers and use a learning strategy that helps them maximize their academic achievement.

Measuring Instructional Effectiveness

A vital component of an effective college or university is a conceptually sound and properly implemented faculty evaluation system. A fair and effective evaluation based on performance and designed to encourage improvement in both the teacher being evaluated and the college or university is a basic need of a good teacher evaluation system (Stronge, 1997). A comprehensive teacher evaluation system should be outcome oriented in that it contributes to the personal goals of the teacher, to the mission of the program, to the school, and to the educational institution, and should provide a fair measure of performance (Stronge, 1997). The evaluation system should be improvement oriented in that it contributes to the personal and professional development needs of the individual teacher as well as improvement within the school (Stronge, 1997). Performance improvement, which reflects the need for professional growth and development of the individual teacher, is improvement oriented (Kyriakides et al., 2006).

When assessment is done properly, it can function as a pivotal component of any formative design for quality (Conrad & Pratt, 1985). Unfortunately, assessment can look threatening and be unpopular with faculty. The assessment process will be more threatening to participants if used for more important decisions (e.g., promotion, tenure, and so forth), thus making it more difficult to get valid and pertinent information (Conrad & Pratt, 1985). The purposes and goals of the assessment process are of critical importance for quality. Questions of purpose and goals left unasked can threaten this quality (Conrad & Pratt, 1985).

There is no single concept of what a teacher should be doing in the classroom, which means there is no single method for evaluating instructional effectiveness. A distinction should be made between summative evaluations and formative evaluations (Algozzine et al., 2004). Summative evaluations are used for administrative decisions such as pay increases, promotion, and tenure. Formative evaluations are used to improve teaching (Spencer & Schmelkin, 2002). Examples of data used for formative evaluations are plans, syllabi, selection of supplemental materials, range and variety of alternative activities, and samples of testing forms and procedures.

Many teachers assert that only colleagues are truly qualified to judge competency (Wilson et al., 1973). Faculty colleagues should only rate the dimensions of teaching that they can observe and are qualified to evaluate (Cohen & McKeachie, 1980), such as course content mastery, course content selection, course organization, appropriateness of objectives, appropriateness of instructional materials, appropriateness of evaluative devices, commitment to teaching, concern for student learning, student achievement based on exam and project performance, and support of departmental instructional efforts. Some teachers also associate the dimensions of research activity and recognition, participation in the academic community, intellectual breadth, relations with students, and concern for teaching with being a good teacher.

Peer nomination and peer ranking methods are typically more summative because they differentiate among faculty on a dimension of overall instructional effectiveness but give no feedback to the teacher to improve his or her teaching (Cohen & McKeachie, 1980). Research on teacher-nominated characteristics suggests that rank, disciplinary area, age, and years of teaching experience are not related to scales of instructional effectiveness (Wilson et al., 1973). Unfortunately, no teacher can be in a position to assess the amount of student learning in a colleague's course without evidence of student achievement, which can be measured in part through student responses on end-of-semester rating forms indicating the students' perceptions of their learning. Also, peer ratings may be based in part on the teacher's reputation generated from student ratings.

Peers can also judge course design and instructional materials. These include course syllabi, reading lists, instructional handouts, and evaluation devices such as examinations and paper assignments. Teachers can fill out a standard form or be interviewed to ascertain information on course goals and objectives and instructional methods used. Examining student papers and projects could provide colleagues with a method of assessing student achievement and cognitive gains in students.

Classroom observations are another method of evaluating instructional effectiveness. Classroom observations can be used both as summative evaluations and formative evaluations. The teacher should be directly involved in determining the time, place and conditions for which the observation should be conducted. Colleagues serving as instructional consultants can be extremely effective if they have the appropriate consultation skills, good observation procedures, and effective feedback delivery procedures (Cohen & McKeachie, 1980).

However, there are many issues with using classroom observations as a form of assessment. First, they are based on the premise that seeing a teacher in action is the best way to gather data for judging that teacher’s effectiveness (Stronge & Ostrander, 1997). This is not necessarily true because classroom visits are typically narrow in scope. Scheduled observations tend to be artificial, each observation has a limited focus, observations are typically infrequent (especially for making generalizations), and one observation cannot evaluate the full repertoire of teacher duties and responsibilities (Stronge & Ostrander, 1997). Personal relationships or alliances between evaluators and their subject can be confounding, especially when evaluators focus attention on their own personal interests and viewpoints. The observation itself also alters the behavior of teacher and students, narrowing the chances of the evaluator seeing a representative sample of teaching (Stronge & Ostrander, 1997). Students may be more qualified to evaluate in-class behavior. Students’ evaluations do not intrude on the class, nor do they create an artificial classroom atmosphere. Therefore, classroom observations should be used in conjunction with students’ evaluations in order to contribute all necessary perspectives to the evaluation process (Cohen & McKeachie, 1980).

Any design for assessment must address methodological approaches to assessing quality of instruction. Student evaluations of teacher performance, teacher evaluations of student performance, and classroom observations tend to emphasize multidimensional perspectives (Conrad & Pratt, 1985). Each method has limitations, so a range of methods should be used. In the end, multiple criteria are necessary and perspectives on teaching must come from multiple sources (Macdonald, 1980b).

Students’ Assessments of Instruction

Background. Students’ assessments of instruction should be one of several sources of information on instructional effectiveness. Students interact with the teacher more than anyone else and are the main source of information about the achievement of educational goals, rapport, degrees of communication, the existence of problems between teachers and students, and the ability of the teacher to motivate (Aleamoni, 1981; Stronge & Ostrander, 1997). They are the only ones who have direct knowledge about teacher classroom practices on a regular basis and should therefore be a key source of data. Students’ evaluations provide a means of communicating between students and teacher. Students can also use these evaluations to obtain information about teachers and courses for course selection, which can indirectly encourage instructional improvement (Aleamoni, 1981).

Students’ evaluations can help describe the learning environment more concisely than other types of measurement. Students are the only participants who are directly and extensively exposed to course elements (e.g., teacher, textbook, homework, course content, method of instruction) and could be considered the most appropriate evaluators of their satisfaction and the quality of those elements (Aleamoni, 1981). The number and content of the evaluation instrument’s dimensions should be based on the purpose of the evaluation. Getting systematic feedback from students, faculty, and administrators about what characteristics are important and what type of feedback is useful is critical for ensuring that the final evaluation instrument is based solely on dimensions of instructional effectiveness believed to be important (Wotruba & Wright, 1975).

Students’ assessment of instruction instruments should not ask students questions that they cannot answer properly. Seldin (1993) believed that some judgments, such as whether the materials used in a course are current and how well a teacher knows the subject matter of the course, are best left up to teachers. Instead, students should be asked to assess their perceptions of their learning in a course and such things as a teacher's ability to communicate at the student's level, rapport with students, ability to stimulate interest in the subject, and ethical and professional behavior in the classroom. Questions can be constructed for the course area, instruction area, and learning area (Aleamoni, 1981). Course area questions should address the course's organization, structure, objectives, difficulty, pace, relevance, content, usefulness, and so on. Questions concerning the instructional area should address teacher characteristics, teacher skill, clarity of presentation, teacher rapport, method of presentation, student interaction, and so on. Questions concerning the learning area should address student satisfaction, student perceived competency, student desire to continue study in the field, and so on.

Regardless of the questions asked and the purposes the students' evaluations will serve, Seldin (1993) said that students' evaluations should contain several open-ended questions to allow students to respond in their own words. These open-ended questions may result in comments that clarify the underlying reasons for particular ratings and point to things that need to be changed. The evaluations can also include space for additional questions selected by the teacher to allow the teachers to shape the form to meet their individual needs as well as those of the department or institution.

Administrative issues. There are many concerns about the appropriateness of students' evaluations. Students' attitudes toward classroom practices can contribute to the teacher's overall effectiveness or can cause disruption. Many feel that students cannot make consistent judgments about the teacher and instruction because of their lack of experience and lack of maturity, and therefore only colleagues are qualified to make these judgments (Aleamoni, 1981). Some research indicates that students' evaluations are subject to some "popularity" pull or "entertainment" pull (Macdonald, 1980b; Stronge & Ostrander, 1997). This creates the belief that student rating forms are unreliable and invalid, yet student evaluations correlate reasonably well with achievement (Macdonald, 1980b). In order for students to make accurate judgments, the rating schemes need to be more than popularity contests. Some teachers even believe that the students should be away for several years from the course, and possibly the university, before they can make accurate judgments (Aleamoni, 1981). Despite all of this controversy, there is a positive relationship between student ratings and learning which can be viewed as a convincing reason to involve students in the process (Stronge & Ostrander, 1997).

A major issue with developing students’ evaluation of instruction instruments is that current instructional methods are far more varied than the items appearing on commonly used ratings instruments can support (Theall & Franklin, 2000). The growth of distance education, particularly asynchronous and on-line teaching and learning, has been a recent issue in evaluation (Theall & Franklin, 2000). The contexts and situations of such courses are substantially different from the traditional face-to-face method of instruction. Technology has the potential to provide powerful teaching and learning tools but is only a passive conductor. A teacher is necessary to construct meaningful experiences and situations and to integrate information, application, analysis, synthesis, evaluation, and reflection (Theall & Franklin, 2001). Using an evaluation designed for face-to-face courses for on-line courses will not address the unique characteristics of these courses and will not provide data specific enough to allow an accurate understanding of the outcomes of instruction (Theall & Franklin, 2000). Different instruments are therefore needed for different methods of instruction.

The quality of the data obtained is determined in part by the method of administering and gathering students’ evaluations. Seldin (1993) believes that evaluations should be administered in the classroom in a formalized manner using a standard set of instructions and giving enough time to complete. If students are permitted to fill the questionnaires out at home and bring them back to class, very few will be returned (Aleamoni, 1981). Students should have all of the necessary materials when completing a questionnaire and should fill it out in their regular classroom near the end of a particular class session. If the teachers are administering their own evaluations, they should read a standard set of instructions and select a student to gather them when completed

(Aleamoni, 1981; Seldin, 1993). This helps ensure that the responses are candid and frank, which is not the case if the teacher will see them at the end of that class period.

Students need to be left with the impression that their frank and honest comments are desired and that their response should not be an attempt to get back at the teacher (Aleamoni, 1981). Telling the students how their ratings will be used helps promote this impression. However, some research suggests that when students are made aware that the purpose of ratings is for tenure and promotion, higher ratings will result (Algozzine, et al., 2004). The students may not respond seriously if they have the impression that the teacher is not really interested in their responses.

Spencer and Schmelkin (2002) found three factors within students’ attitudes towards course and teacher ratings. These are *Reluctance to Do Evaluations*, *Potential Repercussion against Students*, and *Student Opinion Taken Seriously*. The means of *Reluctance to Do Evaluations* and *Potential Repercussion against Students* indicated that students are not reluctant to do evaluations and are not concerned about the potential repercussions against them. Students were skeptical about the use of the ratings as a barometer of student opinion about professors and classes. In other words, students may not pay much attention to evaluations if they are unsure whether their opinions matter and how their ratings are being used. Also, Spencer and Schmelkin (2002) found that the more positively students felt that teaching had affected their lives, the more frequently they felt that teachers were interactive and open to diverse viewpoints, and the more those students thought that their opinions were taken seriously.

If the students believe the teacher is going to see their responses before final grades are reported (and if they are asked to identify themselves on the questionnaire), they will respond more positively and write very few comments (Aleamoni, 1981). Some research has suggested that higher ratings are evident if the students are not anonymous and if the teacher is present when the students are completing the ratings (Algozzine et al., 2004), whereas other research has suggested the opposite (d'Apollonia & Abrami, 1997). Even if this is the case, standardized procedures should still be used for ethical and legal reasons. The rating form should be distributed during the last two weeks of the term and not right before or after final (or other) exams, unless the responses are used solely to improve teaching, where it would be more appropriate to administer the evaluation one-third of the way into the semester to allow the teacher a chance to adjust his or her teaching (Seldin, 1993). Administering the questionnaire immediately before, during, or after the final (or other) examination could result in students responding in an inconsistent manner (Aleamoni, 1981). Some research suggests that student ratings are significantly higher when the evaluation is carried out after the final examination (d'Apollonia & Abrami, 1997). This suggests that students may be rewarding teachers who give them high grades, or that students may be using their grades as one of the indicators of teacher effectiveness. Finally, students should not be allowed to discuss the teacher and course with each other during the evaluation to avoid biases in the results.

Online students' assessments of instruction. Due to the dramatic surge in online computing, the Internet has gained popularity as a collection method for survey information (Carini, Hayek, Kuh, Kennedy, & Ouimet, 2003). Online surveys have many advantages including the reduction in mailing costs, reduced time for implementation, reduced cost in surveying additional respondents, the simultaneous display of response data with the completion of the survey, easier reminders and follow-up with nonrespondents, and ease of the importation of results into data analysis programs (Archer, 2003; Sax, Gilmartin, & Bryant, 2003). There is the potential for more thoughtful responses with online surveys because students are not rushed to complete the evaluation at the very end of class (Thorpe, 2002). Previous studies give some evidence of improved quality and quantity of student open-ended comments. However, there are many disadvantages, such as access issues to the Internet, computer literacy problems, difficulty of sampling e-mail addresses, and differences of display from computer to computer (Archer, 2003).

Irrespective of the survey and the sample selected, it is likely that some members of the sample will not respond to survey questions. Non-response error exists when those who respond to the survey differ on the survey measures from those who do not respond to the survey (Cui, 2003; Sax et al., 2003). This is not synonymous with response bias, which refers to the ways in which the questions themselves are answered. Response bias pertains to respondents answering in socially desirable ways, exaggerating their answers, endorsing items regardless of content, expending little effort in question interpretation and answering, and avoiding extreme response options (Sax et al., 2003).

Additionally, low response rates do not necessarily lead to non-response error (Sax et al., 2003). It is also important to note that the non-response rate alone cannot predict the amount of non-response bias (Cui, 2003; Groves et al., 2006). Despite the possibility that low response rates may not lead to non-response bias, low response rates have always been considered a major problem of survey research. Much research has been conducted on improving response rates through improving survey methods. Some variables found to have a positive effect on response rate include relevance of survey to the respondent, use of a pre-notification letter, use of follow-up letters, the inclusion of incentives with the survey, reactions to the survey sponsor, and shorter survey lengths (Cui, 2003; Groves et al., 2006). Additionally, the "leverage-salience" theory states that the effect of any particular stimulus on a person is a joint function of its centrality to the person (its leverage) and its salience relative to the survey introduction (Groves et al., 2006). In other words, potential respondents can form positive or negative predispositions dependent on what is made salient (e.g., an embarrassing survey topic would stimulate negative predispositions).

There are several ways to increase survey response rate. The most common is the use of incentives, especially prepaid monetary incentives. Research has indicated significantly higher response rates when incentives were used. One study suggested that incentives increase the benefits of participating to those uninterested in the survey topic (Groves et al., 2006). Another study gave evidence that incentives given at the time of the survey will increase response rates (Cui, 2003). There are questions as to whether the use of incentives will bring about a positive response bias; but, currently, no literature has been found to support this assertion. Survey participation can also be increased if the respondent finds that thinking about the topic of interest will be rewarding (Groves et al., 2006). These rewards can be pleasant memories, gratification of knowing that the survey may increase attention to an issue related to self-interests, and so forth. If the topic of the survey is relevant but generates negative thoughts, participation may be suppressed.

There is also research supporting a survey-response hierarchy-of-effects model (Helgeson, Voss, & Terpening, 2002). This model states that the survey-response decision process will follow four steps: *Attention*, *Intention*, *Completion*, and *Return*. *Attention* will be positively related to *Intention*; *Intention* will be positively related to *Completion*; and *Completion* will be positively related to *Return*. In other words, each phase has a significant relationship to the next phase in this process. Under this theory, decisions to complete a survey are not to be viewed as discrete acts, but rather are continuous. Moving respondents through each phase of the process successfully will result in more responses. Each phase can increase the response rate by influencing respondent attitudes and perceptions (Helgeson et al., 2002). Obtaining participation in earlier phases will increase the chances that there will be follow-through in the later phases. There is evidence that incentives can positively affect any phase of the process (Helgeson et al., 2002). Incentives are inherently behavior-modification devices. Also, providing feedback to respondents that helps assuage their curiosity regarding the research outcome can build positive attitudes toward research activities (Helgeson et al., 2002).

The use of paper surveys versus online surveys has been a major concern as it pertains to response rates. Two major questions arise from this concern: whether or not college students will respond to online surveys at higher rate or a lower rate than to paper surveys and whether or not nonrespondents to online surveys differ from nonrespondents to paper surveys (Sax et al., 2003).

Some factors that influence an individual’s decision to complete an online survey include Internet familiarity, ease of completing the survey, privacy and confidentiality concerns, and computer availability (Thorpe, 2002). There is evidence that certain students are more likely to fill out online evaluation of instruction surveys than others, but it is unknown as to whether non-respondent students would also not respond to an in-class paper survey. Previous research suggests that there is no difference in evaluation on instructional responses between the paper method and online method and that non-response bias may not be a concern for online course evaluation methods (Thorpe, 2002).

One study found between-mode variations in responses for the modes of paper-only, paper with web option, web-only with response incentive, and web-only without response incentive (Sax et al., 2003). Most literature suggests that using mixed modes in one administration, such as both paper and online surveys, can be problematic (Sax et al., 2003). Mixed mode administrations are problematic if the results of both modes cannot be equated because of differences in who responds. The lowest response rates were obtained for the two web-only modes; however, the researchers believed these low response rates were more likely attributed to students not regularly checking their email, privacy and confidentiality concerns, and survey length (Sax et al., 2003). Another study suggests that responses in online surveys are more favorable (Carini et al., 2003). Some mechanisms that might contribute to mode differences are “social desirability (responding in socially acceptable ways), acquiescence (the tendency to agree rather than disagree), question order effects (answering later questions to attain consistency with answers to previous questions), and primacy or recency effects (selecting the first or last offered response)” (Carini et al., 2003, pp. 2-3).

There are many ways to increase the response rate for an online survey. Improving the interactive nature of the survey is one way to appeal to respondents. Providing feedback and summary statistics about an individual’s responses can be added incentives to participate. Making the survey more convenient to access and providing computers and Internet to those who do not have computers will also boost response rates. Finally, helping the participants to feel more confident about confidentiality and security will lower suspicions and increase responses. Most studies show higher response rates for paper surveys (Sax et al., 2003); however, some have shown higher response rates for online surveys (Thorpe, 2002).

Usefulness of Students’ Assessments of Instruction

Students’ evaluations are used to provide (a) feedback to faculty about their instructional effectiveness, (b) a measure of instructional effectiveness to be used in personnel decisions, (c) information for students to use in the selection of courses and teachers, and (d) an outcome for research or teaching (Kyriakides et al., 2006; Macdonald, 1980b; Marsh, 1984, 1991, 1992; Marsh et al., 1997; Thorpe, 2002). Despite how students’ evaluations are used, one must keep in mind the various ways they could affect teachers. First, teachers with low ratings might use a retreat-to-basics approach (Greenwald & Gillmore, 1997). This approach involves reducing coverage of course material to increase grades. Instead of using the retreat-to-basics approach, teachers might blame the students and oblige them to work harder by giving weekly paper assignments or quizzes (Greenwald & Gillmore, 1997). For teachers who are already nervous, low ratings could confirm the impression that students are bored or dissatisfied. This would be unlikely to increase teacher motivation and eagerness (McKeachie, 1997). The best way to avoid the implications of low ratings is to give more targeted feedback to the teacher and have the teacher participate in discussions with a consultant or peer. Improved examination performance and affective outcomes as well as higher students’ evaluations can result from feedback with consultation (Marsh & Roche, 1997).

In order to place any credence on students’ evaluations, there needs to be evidence that students are capable of using a reasonably appropriate weighting scheme and have self-insight into how they weight the individual teaching factors (Harrison et al., 2004). Items on students’ evaluations should be important to teaching and must be able to be judged accurately by students (d'Apollonia & Abrami, 1997).

In order for teachers to facilitate self-improvement, they need to gain feedback on their teaching ability. Marsh (2007) stated that from a formative perspective, teachers should be given the most useful feedback about their instructional effectiveness. Marsh (1984) conducted a meta-analysis which suggested that an effective intervention method for instructional effectiveness improvement is feedback combined with a candid discussion with an external consultant.

Administrators are responsible for counseling teachers and for evaluating them with respect to retention, tenure, and promotion. Such decisions include hiring, salary decisions, assigning, reduction in force, performance evaluations, retirement exception, pre-tenure retention/termination, licensing/credentialing, tenure, awards/recognition, post-tenure retention/termination, self-assessment, promotion/career ladder, and mentoring appointments (Wheeler & Scriven, 1997). These administrative decisions could also improve effective teaching through increasing the likelihood that quality teaching will be recognized and rewarded. Tenure and promotion decisions should be based on as many different courses as possible if it is likely that the teacher will teach many different classes during his or her career.

One potential problem is lack of sophistication of personnel committees who use students' evaluations (McKeachie, 1997). Some believe that only simple judgments of instructional effectiveness should be made from students' evaluations. Some faculty members and administrators may have their own stereotypes about what effective teaching involves. Those who do not conform to these stereotypes will be at a disadvantage (McKeachie, 1997). Also, negative information tends to weigh more heavily than positive information in meetings.

There is much debate over whether to use factors, weighted averages, or unweighted averages of students' evaluations and about the merits of each as scores for use in personnel decisions. Some researchers argue that averages are more practical since summative decisions are unidimensional, whereas others argue that individual dimensions should be considered separately (Harrison et al., 2004). A compromise between global ratings and a profile of scores is to use a weighted average of factors (Marsh, 1991; Marsh & Roche, 1997). The weights can be based on the relative importance of each factor as judged by the teacher, the purpose of the students' evaluations, or on the basis of empirical research findings (Marsh, 1991). It has been suggested that there is no difference between weighted averages and unweighted averages (Harrison et al., 2004), but despite this, summarizing students' evaluations to an unweighted average should not be done because different dimensions of students' evaluations will correlate better with different indicators of effective instruction (Marsh, 1984). If students' evaluations are being used solely to provide teachers with formative feedback, averages and overall ratings would be inappropriate (Algozzine et al., 2004; McKeachie, 1997).

Students' evaluations are not only useful for faculty and administrators but also for students. Students seek information that will help them select teachers and courses. Those students who select a class on the basis of information about instructional effectiveness are more satisfied with the quality of teaching than those who indicate other reasons (Marsh, 1984). Information about instructional effectiveness therefore influences course selection.

Research on teaching has the potential to answer questions like how teachers behave, why teachers behave the way they do, and what are the effects of teachers' behaviors. Such research looks at process variables (those on global teaching methods and specific teaching behaviors), presage variables (characteristics of teachers and students), context variables (substantive, physical, and institutional environments), and product variables (student academic/professional achievement, attitudes, and evaluations) (Marsh, 1984).

Research on Psychometric Issues

The summative function of students' assessments of instruction may give an assessment a threatening appearance to faculty, especially without evidence of validity. This threatening appearance comes from concerns about how students' inexperience and biases can affect their perception of their teachers' overall effectiveness. Some teachers feel that students' assessments of instruction may be unreliable and invalid because of a students' lack of experience or immaturity that leads to inconsistent and unfair judgments. Despite these concerns, most research on evaluation instruments suggests that they are indeed reliable and valid in comparison to other measures of effective teaching. In addition, when properly administered and used, students’ assessments of instruction are relatively unaffected by variables believed to be potential biases and provide useful information to faculty, students, and administrators.

Reliability. Reliability is the instrument’s capability of producing stable student responses from one survey administration to another. Reliability is therefore a reflection of the consistency, or degree of agreement, among raters (e.g., students). If the instrument is not reliable, the data and resulting evaluations will be meaningless (Aleamoni, 1981). If students’ evaluations are to be used appropriately and interpreted meaningfully, consistency must be analyzed (Feldman, 1978). Without reliability, the effect of other variables will be masked and there will be added random error. This random error will reduce the power of statistical tests, leading to an increased likelihood that insignificant measures will be found to be significant. This could affect the ability to provide evidence of construct validity.

Marsh’s (1984, 1992) research on students’ evaluations suggests that they are reliable and stable in general. Reliability is important because it assesses agreement among different students within the same class. Reliability addresses the error in the ratings; as error decreases, reliability increases. Schmidt, Viswesvaran, and Ones (2000) state that “reliability is the consistency with which an instrument measures whatever it measures (regardless of the validity of those measurements)” (p. 905). Reliability should be used to justify the use of measuring instruments and has little utility for practical situations besides this (Weiss & Davison, 1981). Unreliable measures mask the effect of the independent variables in the design and reduce the power of any statistical test because they introduce random error. This can lead to conclusions where the researcher is convinced that the measures are significant when in fact they are due to this random error. In order to make inferences about constructs, a researcher needs to have the ability to estimate how much error affects the results. Reliability is therefore necessary in order to provide evidence of construct validity.

Generalizability of ratings across different offerings of the same course by different teachers and across different courses offered by the same teacher is also an important component of reliability. The generalizability across different offerings of the same course with different teachers and across different courses offered by the same teacher will indicate the degree to which students’ responses are a reflection of a teacher and the degree to which they are a reflection of the course. Inter-rater agreement is the most appropriate measure for assessing agreement among different students within the same class. Item analyses, or correlations among responses to different items designed to measure the same component of effective instruction, are also used. Given a sufficient number of students in a class, the reliability of student ratings (based on agreement among all the different students within each class) compares well with the reliability of the best objective tests (.95 for 50 students, .90 for 25 students, .74 for 10 students, and .60 for 5 students; Marsh & Roche, 1997).

Some teachers may be uniquely suited to teach specific courses. Marsh (1984) described several previous studies by other researchers giving evidence that students’ evaluations primarily reflect the effectiveness of the teacher rather than the influence of the course. This can be seen by analyzing the generalizability of ratings across different offerings of the same course by different teachers and across different courses offered by the same teacher. Marsh and Hocevar (1984) demonstrated a consistency of the factor structure of ratings across different sets of courses. Marsh and Roche (1997) produced evidence of this by using correlations between overall ratings of different teachers teaching the same course and for the same teacher teaching different courses. The correlations were low for the former (r = -.05) and much larger for the latter (r = .61), supporting the validity of students’ evaluations more as a measure of instructional effectiveness than as a measure for course effectiveness.

Marsh and Roche (1997) stated that instructional effectiveness can be evaluated by current students, former students, the teacher, the teacher’s colleagues, administrators, and trained observers because there is evidence of students’ evaluations being related to observable teaching behaviors. Some, however, believe that "students cannot recognize effective teaching until after being called upon to apply course materials in further coursework or after graduation", but some cross-sectional and longitudinal studies disagree (Marsh, 1984, p. 717).

This stability of students' evaluations is an important component of reliability. Stability represents whether teacher effectiveness increases, decreases, or remains stable with added experience. The two most common approaches to studying stability are using mean stability, or the stability of means over time, and covariance stability, or the stability of individual differences over time (Marsh, 1992, 2007). Both must be evaluated using longitudinal data in which the same teachers are evaluated on many different occasions over an extended period of time. Marsh (2007) stated that most previous research on stability shows that instructional effectiveness tends to decline with added experience, but that this research has many caveats. The first caveat is that most of this research is cross-sectional instead of longitudinal and therefore cannot evaluate covariance stability and mean stability because of potential selection bias. Cross-sectional studies are poor in predicting what ratings younger teachers will receive later in their careers (Marsh, 1992). The second caveat is that most research is based on studies without any systematic interventions designed to improve instructional effectiveness. Finally, studies of mean stability are typically based on results aggregated across many teachers. When this is done, large individual differences for particular teachers, whether they are declining or improving, is lost.

Marsh's (1992, 2007) research properly addressed these caveats. His conclusion was that university instructional effectiveness based on students' assessments of instruction are highly stable in terms of both mean and covariance stability. Some teachers improved with time while others got worse, but overall there was very little systematic change in instructional effectiveness (Marsh, 1992, 2007). Through the 13 years of his study, Marsh also stated that poor teachers mostly remained poor teachers while good teachers mostly remained good teachers. All of his results pointed to instructional effectiveness being stable, suggesting that teachers do not gain from experience. This could be because university faculty have little or no formal training in teaching or because of a lack of intervention in improving instruction. He suggested that the finding of strong covariance stability facilitated the interpretations of strong mean stability because strong covariance stability implies that the mean stability measures are not random and ratings of the same teacher are highly consistent from one time to another.

Validity. Reliability, content validity, convergent validity, discriminant validity, and criterion-related validity are each components of construct validity (Kraiger & Teachout, 1990). Each of these types of validity should not be equated and each needs to be estimated in order to assess construct validity. Construct validity reflects how useful the evaluation instrument is for measuring the students' view of the effectiveness of instruction. Construct validity can also be defined as the extent to which variability in a measure is a function of the variability of some underlying construct (Kraiger & Teachout, 1990). While construct validity can never be completely present or absent in a validity study, assessing it can provide an understanding of the conceptual framework of the evaluation instrument. Convergent validity is the extent to which alternative methods provide similar information about ratees (e.g., teachers). This asks how well student ratings are correlated with other indicators of effective instruction. Discriminant validity is the extent to which measures of theoretically distinct constructs are empirically unrelated. In the case of student ratings, discriminant validity investigates what factors other than instructional effectiveness influence students' evaluations.

The researcher should specify a conceptual framework about a construct containing the construct's meaning. Criterion-related validity is a reflection of this conceptual framework, or how well the underlying factor structure measures what the instrument intends to measure. The multidimensionality of student ratings makes sense as criterion-related validity because teaching is a complex activity with multiple dimensions. Most teachers will have a systematic profile with some strengths and some weaknesses (Marsh, 1991). An example of this is a teacher who lacks enthusiasm but is organized. Assessments of quality should not be based on a unidimensional concept of quality (Conrad & Pratt, 1985). Every dimension needs recognition in the effort to describe the properties of the process (Conrad & Pratt, 1985). Multidimensionality has a diagnostic utility as teacher feedback and provides a more sophisticated and realistic assessment of the various aspects of teaching (Marsh & Roche, 1997). Also, a well-defined factor structure can provide a safeguard against a halo effect, a generalization from a subjective feeling, external influence, or an idiosyncratic response mode which can affect responses to all items (Marsh, 1984).

Teaching is multifaceted and therefore should be evaluated by measures that reflect that dimensionality. Global or overall ratings are more susceptible to context, mood, and other potential biases than items that are more specific and more closely tied to student ratings (Marsh & Roche, 1997). Items within the same group can be demonstrated to measure separate and distinguishable traits through the use of factor analysis. This shows that interpretation of what is being measured is possible (Marsh, 1984).

Some support for students’ evaluations being multidimensional comes from the nine-factor “Student Evaluation of Educational Quality (SEEQ)” (Marsh, 1984, 1991, 1992; Marsh & Roche, 1997). The nine factors were *Learning/Value*, *Instructor Enthusiasm*, *Organization*, *Individual Rapport*, *Group Interaction*, *Breadth of Coverage*, *Examinations/Grading*, *Assignments/Readings*, and *Workload/Difficulty* (Marsh, 1984, p. 711). Other factors Marsh came across while researching literature on Frey’s Endeavor instrument and the Michigan State Student Instructional Rating System (SIRS) instrument were *Presentation Clarity*, *Personal Attention*, *Class Discussion*, *Student Accomplishments*, *Teacher Involvement*, *Student Interest and Performance*, and *Student-Teacher Interaction*. Feldman (1989) found the four dimensions of *Preparation and Organization*, *Clarity and Understandableness*, *Sensitivity To and Concern with Class Level and Progress*, and *Stimulation of Student Interest* to be of high importance; and, the four dimensions of *Elocutionary Skill*, *Fairness and Impartiality of Evaluation*, *Friendliness and Concern*, and *Respect for Students* to be of moderate importance in discriminating among global ratings received by teachers from their students. Another study suggested the factors of *Research*

Activity and Recognition, *Participation in the Academic Community*, *Intellectual Breadth*, *Relations with Students*, and *Concern for Teaching* as characteristics of effective instruction (Wilson et al., 1973).

Regardless of what factors are underlying the students' evaluation instrument, they should be relatively valid against other indicators of effective instruction. A multi-section validity study of multiple sections of the same course (with the same final exam) being taught by different teachers produced evidence that students' evaluations reflect students' learning (Marsh & Roche, 1997). This evidence came from validity coefficients being higher from some more specific student evaluation components and for multi-item scales than for single items. Marsh, Hau, Chung, and Siu (1997) found that SEEQ responses were valid in relation to student learning, ratings of former students, teacher self-evaluations, and affective course consequences such as plans to pursue further study. Good teachers were rated higher on all SEEQ items and scales than poor teachers (Marsh et al., 1997). Another study found evidence suggesting that courses with higher ratings tended to be courses in which teachers emphasized instructional goals rather than learning facts or concepts, instructional activities other than lectures, and grading methods other than midterms and finals (Franklin & Theall, 1992). Students' evaluations are not perfectly correlated with student learning, but "are the single most valid source of data on teaching effectiveness" (McKeachie, 1997, p. 1219).

Potential biases. Results of research on students' evaluations are sometimes discounted on the grounds that students tend to be too easily swayed by superficialities and are not qualified to evaluate the competency of their teachers (Wilson et al., 1973). These potential biases make up most of the concerns that give rise to the belief that students' evaluations may be unreliable and invalid. However, potential biases may reflect a valid influence if they have a similar influence on multiple indicators of instructional effectiveness (Marsh & Roche, 1997), which would be evidence of convergent validity. Unless potential biases can reduce construct validity, they cannot be described as biasing variables (Algozzine et al., 2004). However, the potential bias may support the construct validity if the pattern of relations between it and multiple dimensions of students' evaluations match *a priori* predictions (Marsh & Roche, 1997). An example of this is having a variable such as class size being correlated with factors such as *Group Interaction* and *Individual Rapport*. This implies that class size actually does affect *Group Interaction* and *Individual Rapport* in a manner that is accurately reflected in ratings and therefore is evidence supporting construct validity. A potential bias must be substantially and causally related to the ratings and relatively unrelated to other indicators of effective teaching in order to constitute a bias to student ratings (Marsh, 1984). There are several methodological problems that need to be overcome in research concerning potential biases, which produce misconceptions about students' evaluations (Marsh & Roche, 1997). Examples of methodological problems include implying causation from correlation, the use of an inappropriate analysis (e.g., using individual students instead of class averages), neglect of the multivariate nature of students' evaluations, neglect of potential biases, inappropriate operational definitions of bias and potential biasing variables, and inappropriate experimental manipulations.

Expected grades make up most of the research on potential biases. If a student has primarily chosen a college to have a good time, easy teachers may be more highly appreciated. On the other hand, if the institution has a higher academic culture, the effect of easy grading may have a negative impact on students' evaluations (McKeachie, 1997). Grade bias is more likely to negatively impact a committee's judgment in either case if the grading pattern is higher than normal (McKeachie, 1997).

There are several major hypotheses concerning expected grades. The first hypothesis, the "grading-leniency hypothesis," proposes that teachers who give higher-than-deserved grades will be rewarded with higher-than-deserved evaluations. The teacher's leniency in assigning grades is expected to influence evaluations instead of the expected grades themselves under this hypothesis. This hypothesis attributes a serious bias to students' evaluations (Greenwald & Gillmore, 1997; Marsh, 1984; Marsh & Roche, 1997). The second hypothesis, the "validity hypothesis," proposes that better expected grades reflect better student learning and that a positive correlation between student learning and evaluations is evidence of validity (Greenwald & Gillmore, 1997; Marsh, 1984; Marsh & Roche, 1997). The third hypothesis, the "students' characteristics hypothesis," proposes that preexisting student variables such as prior subject interest may affect students' learning and grades so that the expected-grade effect on effective instruction is spurious (Marsh, 1984; Marsh & Roche, 1997). The fourth hypothesis is that "students' general motivation" influences both grades and ratings (Greenwald & Gillmore, 1997). Students with high academic motivation should perform better in their classes and should therefore appreciate the efforts of the teacher more fully. The fifth hypothesis is that students' "course specific motivation" influences both grades and ratings (Greenwald & Gillmore, 1997). This hypothesis is built on the notion that a student's motivation can vary from course to course rather than a fixed characteristic of the student. Motivation may or may not be attributed to the instruction. The sixth hypothesis is that students "infer course quality" and their own ability from their received grade (Greenwald & Gillmore, 1997). How people make inferences about their own traits and about the properties of situations in which they act can be described through attribution theories. Favorable outcomes typically lead to the inference that one has desirable traits, whereas unfavorable outcomes lead one to perceive situational obstacles to success. Thus, high grades are likely to be attributed to intelligence or diligence; and, low grades are likely to be attributed to poor instruction. Of these hypotheses, the "validity hypothesis" and the "students' characteristics hypothesis" have the largest body of evidence supporting them (Marsh & Roche, 1997).

In order to get a better handle on the effects of expected-grades, researchers need to develop theoretically defensible operational definitions. Grading-leniency seems to be an attribute of the teacher not individual students within a class. Correlations should therefore be based on class-average results (Marsh & Roche, 1997). If course grade is a reflection of course mastery and achievement, there is support for the validity of the ratings; otherwise, if higher grades reflect easier grading standards, there may be a bias in the ratings (Marsh et al., 1997). Just because there is a relationship between students' expected grade and students' evaluations does not mean that the teacher's grading technique has influenced students' evaluations (Algozzine et al., 2004). This was shown in a path analysis conducted by Marsh (1984), which found that nearly one-third of the expected grade effect could be explained in terms of prior subject interest. Ellis, Burke, Lomire, and McCormack (2003) developed a statistical procedure to derive an adjusted set of ratings of instructional quality that controls for the influence of the average grades given. The procedure involved applying the formula, $\text{Adjusted Rating} = \overline{y} + \left(y - \hat{y}\right)$ to each average teacher rating, where $\overline{y}$ is the average rating given to all of the courses in the sample, y refers to the original unadjusted rating, and $\hat{y}$ is the point on the regression line perpendicular to a given average course grade (i.e., essentially the average teacher rating for teachers with the same average course grade).

Some research suggests that the relationship between students' evaluations and students' expected grades is weak at best (Algozzine et al., 2004). Another study investigated the effect of high grades on course evaluations and students' evaluations and found that the effect of a high grade was positive and applied more to the assessment of the teacher than to the assessment of the course (Ellis et al., 2003). Other studies have suggested that grades correlate positively with students' evaluations (Greenwald & Gillmore, 1997; Marsh et al., 1997). Some researchers proposed that better teachers produce better students with better grades, whereas others stated that it was more likely that giving higher grades to students resulted in a more favorable assessment by those students. The various hypotheses for expected grades and the mixed results indicate that the true nature of expected grades is yet to be determined. Much care should be taken in analyzing and interpreting the effect expected grades may have on students' evaluations.

Validity research has also investigated the effect of student characteristics on students' evaluations. These characteristics include gender, age, year in college, grade point average, academic ability, interest in subject matter, and so on. Marsh, Hau, Chung, and Siu (1997) found evidence that student gender, teacher gender, and their interactions had little to no effect on ratings. Aleamoni (1981) stated that research on gender effects is mixed. Seldin (1993) reported that little or no relationship has been found between students' age, year in college, gender, grade point average, or academic ability and students' evaluations. Whether or not a student was taking the course to fulfill a core requirement of their major was suggested not to have influence on students' evaluations (Aleamoni, 1981). Some research suggests that those taking a course as an elective may rate it higher (Algozzine et al., 2004; Feldman, 1978; Marsh et al., 1997). The belief was that prior subject interest would produce higher ratings.

There is a great deal of research investigating the effect of teacher characteristics (i.e., gender, enthusiasm or expressiveness, experience, research productivity, and rank) on students' evaluations. Teacher's rank, experience, and autonomy have all been suggested to positively influence students' evaluations (Algozzine et al., 2004). Seldin (1993) suggests that no relationship has been uncovered between the teacher's rank, gender, or research productivity and students' evaluations. Aleamoni (1981) stated that research on the effects of the rank of the teacher (e.g., teacher, assistant professor, associate professor, professor) on students' evaluations is mixed. Algozzine et al. (2004) found literature indicating that teacher experience moderates the validity coefficient for student rating and may bias student ratings. He also found literature indicating that the enthusiasm or expressiveness of the teacher had been found to positively influence students' evaluations but had small effects on achievement. Ellis et al. (2003) found that the teacher gender and the number of years taught were not significantly correlated with teacher ratings and course ratings. Teachers who had already taught the course once also tended to receive more favorable students' evaluations (Algozzine et al., 2004).

Courses have characteristics that the teacher cannot control, such as class size, course requirements, course level, subject matter, and topic difficulty. Some research suggests that teachers of large classes will receive lower ratings because students generally prefer small classes

(Aleamoni, 1981; Algozzine et al., 2004; Ellis et al., 2003; Feldman, 1978; Marsh et al., 1997). Teachers of large classes feel an increased challenge, have fewer resources, and may tailor their teaching methods to the size of the course. It is also possible that the course may be large due to the prominent reputation of the teacher. McKeachie (1997) stated that most teachers teach better in smaller classes (i.e., require more papers, encourage more discussion, and are more likely to use essay questions on examinations). Franklin and Theall (1992) found evidence suggesting class size is related to both the instructional choices teachers make and student satisfaction. Low-workload courses may receive higher ratings than high-workload courses (Algozzine et al., 2004; Feldman, 1978, 1989). However, it is important to differentiate between hours spent compensating for poor instruction and work that is constructive in promoting learning and increasing motivation in order to determine the true biasing effect of workload (McKeachie, 1997).

Aleamoni (1981) stated that research on the effects of the level of the course on students’ evaluations is mixed. Some of this research suggests that upper-division courses may be rated higher than lower-division courses (Algozzine et al., 2004; Feldman, 1978, 1989). Another study suggested that course level was not correlated with teacher ratings and course ratings (Ellis et al., 2003). Franklin and Theall (1992) suggest that course level is related to teachers’ choices. It is possible that course level may indirectly contribute to ratings through its association with other intervening variables such as the “electivity” of the course (Feldman, 1978). Students who are required to take a course may rate the teacher lower than those electing to take a course (Aleamoni, 1981). It is also believed that the subject matter of the course has possible effects on students’ evaluations. Higher ratings are believed to be related to humanities, fine arts, and languages teachers as opposed to social science, physical science, mathematics, and engineering teachers, but there is limited systematic research to validate this (Feldman, 1978). Other research indicates that there is no evidence of time and day of the course having an influence on students’ evaluations (Aleamoni, 1981; Feldman, 1978).

Students' experience with future teachers may be affected by experiences with their current teachers (Algozzine et al., 2004). This is called the "halo effect." The attitude can be carried to other teacher evaluations. When students feel that teachers do not care about their learning, a negative "halo effect" may result, as opposed to students who feel that teachers do care about learning, resulting in a positive halo (Algozzine et al., 2004; McKeachie, 1997). Alternatively, Marsh and Hocevar (1984) demonstrated a lack of the halo effect in student ratings.

Statement of the Problem

There are two purposes of the Students' Assessment of Instruction (SAI) instrument at Western Carolina University (WCU). The first purpose is to obtain summative information on which to base personnel decisions (e.g., tenure, promotion, reappointment, and merit pay) by administering the appropriate forms of the SAI across the university. For this purpose, colleagues' reviews of teaching (e.g., classroom observations and/or reviews of teaching materials) and teachers' self-reports and evaluations are used in conjunction with the SAI instrument. In this respect, WCU recognizes that no single source can provide sufficient information to make a valid judgment about overall teaching effectiveness. The second purpose is to obtain formative information to use as a basis for making decisions for improving instruction. The summative purpose of the SAI could be formative in that it will provide a basis for informed administrative decisions leading to an increase in the likelihood that good teachers will be reappointed, receive tenure, and so on. The SAI instrument was designed based on the five factors of *Organization and Clarity*, *Enthusiasm and Intellectual Stimulation*, *Rapport and Respect*, *Feedback and Accessibility*, and *Student Perceptions of Learning*. It is also believed that these five factors can be summarized into a higher-order factor labeled *Perceived Instructional Effectiveness*. This higher-order factor represents overall instructional effectiveness as perceived by the students.

Hypothesis 1: The factor structure of the SAI will fit these five first-order factors and the second-order factor, and no other factor structure will explain the SAI better than this. All factor loadings for the first-order factors and second-order factor will be positive.

Hypothesis 2: The reliability of this model will be adequate for its intended use.

The current study will investigate discriminant validity to determine if potential biases are affecting the ratings. The biggest potential bias is expected grade. The literature has revealed several hypotheses on the effect of expected grades. These are the "validity hypothesis" (better expected grades reflect better learning by students and therefore supports the validity of students' evaluations), "students' characteristics hypothesis" (preexisting student characteristics may affect students' learning and grades so that the expected grade effect is spurious), "grading-leniency hypothesis" (teachers who give higher-than-deserved grades will be rewarded with higher-than-deserved evaluations), "student motivation hypothesis" (students' general motivation influences both grades and ratings), "course specific motivation hypothesis" (students' motivation can vary from course to course rather than being a fixed characteristic of the student and can influence both grades and ratings), and the "infer course quality hypothesis" (students infer course quality and own ability from their received grade). The current study will attempt to determine which of these hypotheses is supported.

Hypothesis 3: The "validity hypothesis" will explain most of the relationship between expected grade and ratings (convergent validity) and other hypotheses will not be significant enough to suggest that expected grades invalidate the SAI.

Other potential biases that need to be investigated are attempted hours, reason for taking the course, level of prior interest, course difficulty, course subject, amount of reading, amount of writing, overall workload, pace, hours per week required outside of class, teacher gender, teacher rank, class size, and course level. It is important to note that insignificant associations between any of the potential biases and the SAI factors are not evidence against reliability and validity because this would imply that the potential biases do not influence the SAI factors at all in any way. Hypotheses need to be tested for each of these potential biases:

Hypothesis 4a: Attempted hours, reason for taking the course, level of prior interest, student gender, teacher gender, teacher ethnicity, teacher age, class size, course level, course subject (defined by creating groups based off of 2-digit CIP codes), pace, and course subject will not be associated with *Perceived Instructional Effectiveness*.

Hypothesis 4b: Course difficulty will be positively associated with the *Student Perceptions of Learning* factor illustrating that course difficulty promotes learning and *Rapport and Respect*, implying that students respect teachers who challenge them. A negative association between course difficulty and *Perceived Instructional Effectiveness* would suggest that easier courses receive higher ratings. A non-significant association would suggest that course difficulty offers no evidence for or against construct validity.

Hypothesis 4c: The amount of reading, the amount of writing, and overall workload will be positively associated with the *Student Perceptions of Learning* factor and the *Feedback and Accessibility* factor implying that a higher workload is appropriate and fair for the course and that all learning is not occurring inside of the classroom. A negative association *with Rapport and Respect* would imply that students do not respect teachers who give heavy workloads and that a significant amount of learning may be occurring outside of the classroom. No significant associations would imply that these potential biases offer no evidence for or against construct validity.

Hypothesis 4d: Pace will be positively associated with *Enthusiasm and Intellectual Stimulation* suggesting that courses are faster as a result of this factor. A non-significant association would imply that pace offers no evidence for or against construct validity.

Hypothesis 4e: The number of hours per week required outside of class will be negatively associated with the factors of *Student Perceptions of Learning* and *Rapport and Respect* indicating that the workload outside of class is fair and appropriate and that a significant amount of learning does not occur outside the classroom through assignments. A non-significant association would imply that the number of hours per week required outside of class offers no evidence for or against construct validity.

Hypothesis 4f: Positive associations are expected between *Perceived Instructional Effectiveness* and teacher rank implying that teachers who have been teaching longer and who have more credentials are better teachers. A non-significant association would imply that teacher rank offers no evidence for or against construct validity.

METHOD

Participants

The participants were volunteers from the entire population of Western Carolina University students for the fall 2007 and spring 2008 semesters. For the spring 2007 semester, participants were students taking courses in the departments of Applied Criminology, Chemistry and Physics, Marketing and Business Law, Political Science and Public Affairs, and Psychology. The population consisted of both undergraduate and graduate students.

Materials

The instrument used was the Student Assessment of Instruction (SAI), developed by a committee of faculty members at Western Carolina University (WCU). Different versions of the SAI were created for each instructional method. The versions that were used in this study were the Activities Course Form (for courses with a substantial component of physical activity directed at the learning of skills), Independent Research Course Form (for project courses and theses), Internship/Practica/Clinical Course Form (courses where the teacher is a supervisor and may have more contact with an off-campus supervisor than directly with students), Laboratory Course Form (for lab-based courses involving active work by the student under supervision), Online Course Form (for courses that are largely delivered over the web in an asynchronous manner), Seminar Course Form (for small classes designed to engage students in frequent participation), Standard Course Form (for the most common type of course which is likely to include some mix of lecture, discussion, in-class activities, etc.), and the Studio-Performance Course Form (for courses that would involve a large amount of one on one instruction that may occur in relatively unstructured settings and in which the teacher's role is largely to provide feedback rather than direct instruction; see Appendix A). The SAI contains a total of 20 questions, four questions for each of the factors of *Organization and Clarity*, *Enthusiasm and Intellectual Stimulation*, *Rapport and Respect*, *Feedback and Accessibility*, and *Student Perceptions of Learning*. These questions are 5-

point Likert-type scales for the spring 2007 and fall 2007 semesters (Strongly Agree; Agree; Neutral; Disagree; Strongly Disagree). In addition to the 20 Likert-type questions, there are open-ended questions (see Appendix A). The fall 2007 instrument also contained 11 Liberal Studies questions. A policy was added after the fall 2007 semester stating that no questions or items regarding program evaluation or any other objectives were to be included in the SAI. For the fall 2007 semester there were 11 additional validity questions asking about student characteristics (see Appendix B). The spring 2008 instrument was identical to this, except that it utilized a 4-point Likert type scale instead of a 5-point (the "Neutral" response category was removed), had a "Not Applicable" response category, and did not have the 11 Liberal Studies questions and 11 validity questions.

Procedure

The SAI instrument was piloted in the spring 2007 semester for the departments of Applied Criminology, Chemistry and Physics, Marketing and Business Law, Political Science and Public Affairs, and Psychology. Following this pilot, the instrument was administered for all courses in all departments. The instrument was administered online through CoursEval®, a third party vendor application. It was the responsibility of each faculty member to review the tool and select the appropriate course form for his/her individual course. All courses were evaluated, including low enrolled courses containing fewer than 5 students. These low enrolled courses included a disclaimer stating that the teacher may be able to determine from whom the comments came because of the class size. No teacher was allowed to see their ratings and comments until after final grades were posted. All evaluations were strictly confidential. The SAI was also administered to non-full semester courses. The policy was to open the SAI no later than when 80% of the class meetings have been completed and to close them no later than when 90% of the class meetings have been completed (excluding the final examination period). All evaluations had to be open for a period no less than one week.

The current analysis investigated the reliability, validity, and factor structure of spring 2007, fall 2007, and spring 2008 SAI data from the Activities Course Form, Independent Research Course Form, Internship/Practica/Clinical Course Form, Laboratory Course Form, Online Course Form, Seminar Course Form, Standard Lecture Course Form, and the Studio-Performance Course Form. The Standard Lecture Course Form is investigated first since most courses are of this instructional method, thereby making this course form the most important. The Seminar Course Form and Online Course Form are investigated after the Standard Lecture Course Form because they are second in importance. All other course forms are investigated after these. Student participation was confidential and voluntary. The validity questions were added to all SAIs only for the fall 2007 semester.

Confirmatory factor analysis (CFA) was used to test the theory of the construct and analyze the model's fit. Also, the error variances of the items reflect score unreliability if the model is specified correctly (Thompson, 2004). The current study undertook a CFA, using software package EQS version 6.1, in order to determine if the proposed first-order factors of *Organization and Clarity*, *Enthusiasm and Intellectual Stimulation*, *Rapport and Respect*, *Feedback and Accessibility*, and *Student Perceptions of Learning* and proposed second-order factor of *Perceived Instructional Effectiveness* are the best fit for the SAI instrument. This analysis was used to analyze hypothesis 1. Some of the CFAs were hierarchical, which is predicated on the assumption of a multidimensional construct with a well-defined set of first-order factors.

Four models were tested for each semester and each course form to determine if the hypothesized theory has an adequate fit and explains the underlying theory of the construct the best (see Appendix C). This is because testing multiple plausible models is desirable (Thompson, 2004). Each model was fit using maximum likelihood estimation on the covariance structure for only those observations that were complete. The first model was a one-factor model. This model suggested that all 20 questions can be explained by just one factor. In part, this model was included because it was suggested by some departments who conducted a preliminary exploratory factor analysis of their own that only one factor was extracted. One of the ultimate goals of exploratory factor analysis is to choose the number of factors by identifying a residual factor that has no practical, psychological, or statistical significance (Weiss, 1971). Many researchers extract all factors having eigenvalues greater than one (Abrami & d'Apollonia, 1991; d'Apollonia & Abrami, 1997) while others use Cattell's scree test. Unfortunately, it is impossible to identify the residual factor correctly using these methods. Also, eigenvalues only contain insights about the unrotated factors and the amount of information contained by them (Thompson, 2004). Rotating factors redistributes the variance making it more difficult to determine the number of factors in an exploratory factor analysis. The rotated factor structure is also more likely to be representative of other samples in the same population (Weiss, 1971). Therefore, it was expected that the one-factor model would be the least adequate model. The second model was a five-factor orthogonal model where the five factors of *Organization and Clarity*, *Enthusiasm and Intellectual Stimulation*, *Rapport and Respect*, *Feedback and Accessibility*, and *Student Perceptions of Learning* were not correlated with each other and can explain each course's form. The third model was identical to the second except that the five factors were allowed to correlate with each other, producing an oblique model. The fourth and final model was the hypothesized model. This model was a higher-order model, the first-order factors being the same as those in the third model. The second level contained only one factor which was predicted to be an overall measure of students' assessment of instruction. First-order variables are narrow in scope whereas the higher-order variables are broader in scope, so this second-order factor was theorized to be a summary of all five first-order factors. Higher-order factors can be interpreted based on their relationships with the original factors, just like first-order factors. Higher-order factors are also measured with the same degree of accuracy as first-order factors and should not be ignored unless they account for a small percentage of the variance, such as 2-3% (Thompson, 2004). The results of the hypothesized model's CFA were then used to compute factor scores by using the magnitude of each factor loading to weight each variable.

For each of these models, a choice was specified as to whether to constrain any unstandardized factor structure coefficient or to constrain the factor variances. This choice is necessary for standardizing and giving the factors a scale of measurement, and the number 1.00 is typically used in either case (MacCallum, 1995). Constraining the unstandardized factor structure coefficients allows the factor variances to be measured as some function of the constrained variable. However, constraining the factor variances is less restrictive. This method presumes that the parameters are independently estimated for different groups and that the same model fits different groups (Thompson, 2004). This typically results in better model fit. Also, constraining the factor variances each to 1.00 allows us to interpret the covariances of the factors as factor correlations and to compare the structure coefficients with each other since they are all freed (Thompson, 2004). This interpretation of the structure coefficients can be used to determine if there are any items within a given factor that explain less of that factor than the other items within that factor. These items will also have a higher variance, and removing them might improve model fit.

Several model fit statistics were included in the analyses. The first was the normed fit index (NFI), which compares the χ^2 between the tested model and the baseline model, presuming that the measured variables are completely independent. A NFI of .95 signifies excellent model fit (Thompson, 2004). The second model fit statistic was the comparative fit index (CFI). This index assesses model fit relative to a null, or independence, model. Similar to the NFI, a CFI of .95 also signifies excellent model fit (Hu & Bentler, 1999; Thompson, 2004). Finally, the root-mean-square error of approximation (RMSEA) statistic was included. This statistic estimates how well the population covariances can be reproduced from the model parameters. A RMSEA of .06 or less signifies excellent model fit and a RMSEA of 0 indicates that the estimated model reproduces the population covariances exactly (Hu & Bentler, 1999; Thompson, 2004). The CFI and RMSEA model fit statistics are two of the most sensitive to misspecified factor loadings (Hu & Bentler, 1999).

After hypothesis 1 is addressed reliability (hypothesis 2) will be analyzed. Classical test theory methods have traditionally been used to analyze reliability. The primary goal of classical test theory is to use reliability coefficients and standard errors to evaluate the quality of observed test scores. Raters' (or students') observed scores from an instrument are used to estimate the true score of the ratee (or teacher). The average instability of using observed scores as an indication of true scores for a group of ratees is the standard error (Suen & Lei, 2007). In other words, the observed score minus the standard error should equal the true score. Standard errors could be impacted by the number of raters, how many items the instrument has, time, setting, and so on. Lower standard errors result in higher reliability coefficients. Classical test theory uses different strategies to assess reliability such as the test-retest, inter-rater, and internal consistency methods (Suen & Lei, 2007). Each method has a different type of reliability coefficient, which has a different error. All of these errors need to be considered simultaneously because the observed score is the product of all of them, but unfortunately classical test theory contains no mechanism for combining these errors (Suen & Lei, 2007). Also, these errors can be characterized as random or fixed. An example of an error that is random is when the actual raters are a sample of all possible raters. An example of an error that is fixed is when the actual raters constitute all possible raters. Classical test theory can only analyze errors that are fixed and therefore will have a tendency to over-represent the reliability of the scores (Suen & Lei, 2007). When all sources of error are considered together or there are sources of error that are random instead of fixed, the reliability will be lower.

Generalizability theory (g-theory) was introduced in 1963 by Cronbach, Gleser, and Rajaratnam as an extension to classical test theory, to account for the flaws of classical test theory and to provide a better measure of reliability and validity (see Cronbach, 2004). Cronbach (2004) endorsed g-theory and has stated that alpha "coefficients are a crude device that does not bring to the surface many subletics implied by variance components" (p. 394). G-theory provides a mechanism for combining all sources of error and their interactions simultaneously in order to achieve the most unbiased stability coefficient value and also allows for the analysis of sources of error that are random in nature (Suen & Lei, 2007).

G-theory will therefore be used to analyze the accuracy of generalizing from a teacher's evaluation for a course to the average score that teacher would have received under all other possible conditions. Some of these conditions could be a different semester, different group of students, using different items on the SAI, and so forth. G-theory was also used to produce decision studies containing generalizability coefficients (g coefficients), similar to intraclass correlation coefficients (Shavelson & Webb, 1991). An intraclass correlation coefficient is a measure of inter-rater reliability. It assesses rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects. This g coefficient is the ratio of the universe-score (i.e., the mean of the observations over the universe of generalization) variance to the expected observed-score variance and is the accuracy of generalization from an observed score to the universe score. The universe are the conditions of observation over which will be generalized, that is, a reflection of the construct under study (Smith, 1979). Smith stated that "generalizability theory unites both reliability and validity theory in that the generalizability coefficient indicates the reliability of the procedure but rests upon the assumption that one can validly interpret a measure as representative of a specified construct" (1979, p. 80). Correctly specifying the universe is therefore critical if specific decisions are going to be made from the SAI. More important than this g coefficient is g-theory's ability to estimate the variance components for each source of error. These sources of error are typically referred to as facets. G-theory can handle complex, multifaceted designs containing crossed facets, nested facets, random facets, and fixed facets.

The data in the current study contains random facets. G-theory will then be used to assess hypothesis 2 and to ensure that the reliability coefficients are not overestimated. The generalizability designs that were analyzed were (*x* = crossed facets, : = nested facets, s = student, c_1 = course, c_2 = course section, t = teacher, i = item):

$$s : c_2 : t \times i \tag{1}$$

$$s : t : c_1 \times i \tag{2}$$

In other words, the first design can be expressed as students being nested within course sections which are nested within-teachers, and that the same set of items is administered to all students. Higher generalizability with respect to teachers' instructional ability is crucial if personnel decisions are going to be made based on students' evaluations (Smith, 1979). Thus, the first design defines the teacher as the universe and the second design defines the course as the universe. Since one teacher may teacher multiple courses and multiple sections of the same course (each course section is essentially a different condition of the teacher), the course section was used in the first model. The second model used the course instead of individual course sections because there is usually only one teacher teaching a course section (each teacher is essentially a different condition of the course), and because possible decisions that can be made from this model involve judgments on courses, programs, and curriculums (Smith, 1979). Since the student, course section, and teacher facets are nested, it will not be possible to analyze these facets individually; instead they will show up in the analysis as interaction terms. It will also be impossible to determine from the g-theory analysis alone if the non-universe facets are errors or evidence of validity. This will have to be analyzed using hierarchical linear modeling (HLM). These g-theory designs assess the SAI's dependability for making judgments about teachers' instructional ability and courses, respectively. If the first design yields a higher variance for teachers than the variance for courses from the second design then this would be evidence of reliability and validity. Also, g coefficients for the decision studies were computed using the formulas $\varepsilon\rho^2(C,S,I)$ for the $student : course\ \sec tion : teacher \times item$ design, which generalizes over courses, students, and items, and $\varepsilon\rho^2(T,S,I)$ for the

$student : teacher : course \times item$ design, which generalizes over teachers, students, and items (see Equations 3 and 4). The g coefficients, along with whether or not more variability in ratings is explained by teachers instead of courses, will indicate how reliable each course form is. The g-theory analysis will be conducted using software package SAS version 9.1.3 using the MIXED procedure with MIVQUE0 (minimum variance quadratic unbiased estimation of the covariance parameters) method estimation.

$$\varepsilon\rho^2(C,S,I) = \frac{\sigma^2(t)}{\sigma^2(t) + \frac{\sigma^2(c:t)}{n'_c} + \frac{\sigma^2(s:c:t)}{n'_c n'_s} + \frac{\sigma^2(t \times i)}{n'_i} + \frac{\sigma^2(c \times i:t)}{n'_c n'_i} + \frac{\sigma^2(s \times i:c:t)}{n'_c n'_s n'_i}} \qquad (3)$$

$$\varepsilon\rho^2(T,S,I) = \frac{\sigma^2(c)}{\sigma^2(c) + \frac{\sigma^2(t:c)}{n'_t} + \frac{\sigma^2(s:t:c)}{n'_t n'_s} + \frac{\sigma^2(c \times i)}{n'_i} + \frac{\sigma^2(t \times i:c)}{n'_t n'_i} + \frac{\sigma^2(s \times i:t:c)}{n'_t n'_s n'_i}} \qquad (4)$$

As long as teachers think potential biases affect student ratings, they will think student ratings are invalid. Analyzing the relationship between potential biases and ratings is necessary for addressing this and providing evidence of convergent validity and discriminant validity. Understanding these relationships is important because construct validity can never be completely present or absent (Marsh, 1984). A potential bias must be causally related to the ratings and unrelated to other indicators of effective teaching in order to be a bias which also makes the understanding of relationships important. Hypotheses 3 through 4a-f will therefore be analyzed next, using software package HLM version 6.06.

Of all potential biases, expected grade has received the most attention. All expected grade hypotheses need to be evaluated carefully in order to determine the influence of expected grade. Expected grade, overall GPA, actual class average grade, attempted hours, course difficulty relative to other courses, the interaction between interest level in the course and expected grade, the interaction between overall GPA and actual class average grade, the interaction between expected grade and overall GPA, and the interaction between expected grade and attempted hours were used for this purpose. Evidence supporting the "validity hypothesis" would be finding a positive association between expected grade and ratings. Unfortunately, this association does not mean that this is purely due to better teachers producing better students with better grades. This result may be due in part to students giving a more favorable assessment when receiving higher grades. Evidence supporting the "student general motivation hypothesis" would be finding a positive association between ratings and the interaction between expected grade and overall GPA and/or between ratings and the interaction between expected grade and attempted hours. Evidence supporting the "course specific motivation hypothesis" would be a positive association between course difficulty relative to other courses and/or the interest level in the course and ratings. Evidence for the "grading-leniency hypothesis" would be a significant association between ratings and the interaction between overall GPA and actual class average grade. A significant association between ratings and the interaction between expected grade and interest level in the course would be evidence for the "students' characteristics hypothesis." The "infer course quality hypothesis" cannot be investigated since it relies on students' actual grades in the course, which are not in the data sets.

The levels were similar to that of the g-theory analysis, with student characteristics (e.g., perceived amount of reading, perceived amount of writing, perceived workload, perceived course pace, gender, reason taking course, interest level, perceived course difficulty, attempted hours, hours spent outside of the classroom, expected grade, and overall GPA) making up level 1, class section characteristics (e.g., course subject, course level, course credit hours, whether or not the course can be used for liberal studies, course enrollment, and course average grade) making up level 2, and faculty characteristics (e.g., ethnicity, age, gender, and rank) making up level 3. The HLM models had to be built from the bottom level up. Each model will have the fall 2007 *Perceived Instructional Effectiveness* structure coefficient as the outcome variable. The first model was a one-way random effects base model which incorporated all three levels. This model is fully unconditional because no predictors were specified at any of the three levels. The primary purpose of this model is to disentangle how much student-level variance for *Perceived Instructional Effectiveness* is attributable to within-section variance, between-section variance, within-teacher variance, and between-teacher variance. This allows for the computation of reliability estimates for the class level and the teacher level. These reliability estimates are measures of the true score relative to the observed score, and are similar to the g coefficients in the g-theory analysis in that they are interrater reliabilities within their level.

The second model that was estimated was a full Level-I random coefficients model utilizing student-level characteristics and potential biases to predict *Perceived Instructional Effectiveness*. The Level-II and Level-III intercepts were random, allowing them to vary across course sections and teachers. The student (within) level variance that was explained by the Level-I variables was calculated for this model. The third model estimated was the Level-II model which had the intent of explaining the unexplained variance due to between-section differences. The intercepts for Level-II and Level-III were again allowed to vary. This model included all the Level-II course section variables, some specifically due to their hypothesized interaction with Level-I variables. The proportion of variance explained by the Level-II variables was calculated for this model. Finally, the last model estimated was a level-III model with all the faculty variables. The proportion of variance explained by these additional variables was again estimated. This model contained variables from all three levels and was used to explain how much variance each level contributes to *Perceived Instructional Effectiveness*. This model was then used to assess hypotheses 3, 4a, 4b, and 4f. This final model was then estimated using some of the first-order factors as the outcome to assess hypotheses 4b-e. Below are notations for a 3-level hierarchical linear model.

$$Y_{ijk} = \pi_{0jk} + \sum_{p=1}^{P} \pi_{pjk} a_{pjk} + e_{ijk} \qquad \text{(level 1 model)}$$

$$\pi_{pjk} = \beta_{p0k} + \sum_{q=1}^{Q} \beta_{pqk} X_{qjk} + r_{pjk} \qquad \text{(level 2 model)}$$

$$\beta_{pjk} = \gamma_{pq0} + \sum_{s=1}^{S} \gamma_{pqs} W_{sk} + \mu_{pqk} \qquad \text{(level 3 model)}$$

where $\pi_{pjk}\left(p = 0,1,\ldots,P\right)$ are level-1 coefficients;
$\beta_{pqk}\left(q = 0,1,\ldots,Q\right)$ are level-2 coefficients;
$\gamma_{pqs}\left(s = 0,1,\ldots,S\right)$ are level-3 coefficients;
a_{pjk} is a level-1 predictor;
e_{ijk} is a level-1 random effect;
X_{qjk} is a level-2 predictor;
r_{pjk} is a level-2 random effect;
W_{sk} is a level-3 predictor; and
μ_{pqk} is a level-3 random effect

G-theory and HLM were chosen as the statistical techniques to analyze hypotheses 2, 3, and 4a-f because there is great controversy when analyzing reliability and validity about whether to use students or class-average ratings as the unit of analysis. For example, many believe class-average responses are appropriate since they are not affected by students' implicit theories about dimensions of teacher behaviors. Larson, however, suggests that these implicit theories can generalize across students, thus affecting class-average responses (see Marsh, 1984). Averaging responses can also mask the systematic variance in individual student ratings. If different students view instructional effectiveness differently, this would be lost with class-average responses.

This unit of analysis problem can lead to mis-estimations. Disaggregating higher order variables, such as course and faculty characteristics, to the individual level (i.e., the student) can violate the statistical assumption that observations are independent of one another that is common for most procedures (Ethington, 1997). By disaggregating, the standard errors may be underestimated leading to rejecting hypotheses that should not be rejected (Patrick, 2001). Higher order variables may not impact the individual-level the same way, making disaggregation a poor method for analysis (Ethington, 1997). Aggregating individual characteristics (i.e., the student) to higher order variables (i.e., the course) is also a problem. Aggregating individual level data does not account for within-group variability which often accounts for the majority (80-90%) of total variation (Ethington, 1997). When student ratings are aggregated to the classroom level, the results will depend upon the degree of variation from classroom to classroom. This aggregation assumes a high inter-rater agreement of the students and is strongly influenced by the number of students per class (Miller & Murdock, 2007). It is possible for the size and direction of correlations to be different between the two units of analysis.

Traditional statistics, such as Cronbach's alpha, do not address the nested structure of our data (Miller & Murdock, 2007). This nesting structure is students grouped within classes grouped within teachers. Accounting for this nesting structure is critical because students within classes should be more similar to one another than those in different classes at the student level (Miller & Murdock, 2007). Students in the same class may have a common set of experiences resulting in levels of interdependence. At the class level, reliability and validity would be representative of the extent that the students express similar perceptions of their classroom environment (Miller & Murdock, 2007). In other words, the student level measures variation among students within the class and around the class's "true score" (e.g., true perception of the teacher), the class level measures variation across classrooms in which the students are nested in terms of how different classes vary around a teacher, and the teacher level measures variation across teachers in which classes are nested.

RESULTS

Standard Lecture Course Form

The Standard Lecture Course Form is the course form with the most data since most courses are taught with this instructional method, making this course form the most important. Model fit for the second-order model was excellent for each semester for the Standard Lecture Course Form (see Table 1). The NFI and CFI values met the ideal cutoff.

All questions loaded positively and significantly at the alpha = .05 level in their expected first-order factor (see Table 2). This is also true for the first-order factor loadings on *Perceived Instructional Effectiveness* (see Table 3). Therefore, it appears that this instrument does not need any changes. Overall, the confirmatory factor analysis (CFA) yields strong evidence of construct validity.

Table 1

Standard Lecture Course Form Goodness-of-fit Indices

Model	N	NPAR	NFI	CFI	RMSEA
Spring 2007					
One-Factor	1793	21	0.899	0.903	0.114
Five-Factor Orthogonal	1793	25	0.707	0.710	0.197
Five-Factor Oblique	1793	25	0.966	0.970	0.066
Second-Order	1793	26	0.959	0.963	0.073
Fall 2007					
Model	N	NPAR	NFI	CFI	RMSEA
One-Factor	9603	21	0.892	0.893	0.118
Five-Factor Orthogonal	9603	25	0.713	0.714	0.193
Five-Factor Oblique	9603	25	0.965	0.966	0.069
Second-Order	9603	26	0.961	0.962	0.073
Spring 2008					
Model	N	NPAR	NFI	CFI	RMSEA
One-Factor	8687	21	0.892	0.893	0.116
Five-Factor Orthogonal	8687	25	0.713	0.713	0.189
Five-Factor Oblique	8687	25	0.968	0.969	0.064
Second-Order	8687	26	0.964	0.964	0.069

Note. All chi-square values are significant at α = .001.

Table 2

Standard Lecture Course Form Pattern and Structure Coefficients

Spring 2007			
		Pattern	Structure
Factor	Question	Coefficient	Coefficient
Organization and Clarity	1	0.475	0.816
	2	0.643	0.901
	3	0.582	0.891
	4	0.574	0.894
Enthusiasm and Intellectual Stimulation	5	0.433	0.810
	6	0.615	0.890
	7	0.615	0.917
	8	0.600	0.892
Rapport and Respect	9	0.653	0.912
	10	0.561	0.816
	11	0.567	0.741
	12	0.539	0.803
Feedback and Accessibility	13	0.629	0.858
	14	0.549	0.859
	15	0.576	0.845
	16	0.628	0.802
Student Perceptions of Learning	17	0.582	0.939
	18	0.577	0.947
	19	0.575	0.929
	20	0.537	0.899
Fall 2007			
		Pattern	Structure
Factor	Question	Coefficient	Coefficient
Organization and Clarity	1	0.448	0.802
	2	0.617	0.900
	3	0.553	0.887
	4	0.555	0.882
Enthusiasm and Intellectual Stimulation	5	0.374	0.762
	6	0.591	0.890
	7	0.583	0.924
	8	0.566	0.903
Rapport and Respect	9	0.622	0.912
	10	0.524	0.789
	11	0.505	0.695
	12	0.524	0.795
Feedback and Accessibility	13	0.577	0.866
	14	0.531	0.869
	15	0.533	0.858
	16	0.573	0.807
Student Perceptions of Learning	17	0.544	0.932
	18	0.551	0.948
	19	0.554	0.936
	20	0.505	0.886

Table 2 (Continued)

Spring 2008			
Factor	Question	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1	0.345	0.784
	2	0.478	0.883
	3	0.443	0.874
	4	0.436	0.877
Enthusiasm and Intellectual Stimulation	5	0.313	0.747
	6	0.485	0.874
	7	0.480	0.907
	8	0.465	0.889
Rapport and Respect	9	0.494	0.904
	10	0.433	0.809
	11	0.407	0.699
	12	0.412	0.790
Feedback and Accessibility	13	0.474	0.862
	14	0.436	0.873
	15	0.446	0.867
	16	0.470	0.812
Student Perceptions of Learning	17	0.438	0.921
	18	0.443	0.939
	19	0.450	0.934
	20	0.402	0.884

Note. All coefficients are significant at the α = .05 level.

Table 3

Standard Lecture Course Form Pattern and Structure Coefficients for Perceived Instructional Effectiveness

Spring 2007		
First-Order Factor	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1.481	0.959
Enthusiasm and Intellectual Stimulation	1.572	0.947
Rapport and Respect	1.305	0.948
Feedback and Accessibility	1.367	0.932
Student Perceptions of Learning	1.551	0.953
Fall 2007		
First-Order Factor	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1.415	0.944
Enthusiasm and Intellectual Stimulation	1.539	0.945
Rapport and Respect	1.290	0.947
Feedback and Accessibility	1.362	0.922
Student Perceptions of Learning	1.513	0.952
Spring 2008		
First-Order Factor	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1.336	0.949
Enthusiasm and Intellectual Stimulation	1.386	0.944
Rapport and Respect	1.219	0.945
Feedback and Accessibility	1.280	0.917
Student Perceptions of Learning	1.415	0.942

Note. All coefficients are significant at the $\alpha = .05$ level.

The generalizability theory (g-theory) analysis also gives support for using the Standard Lecture Course Form. This analysis had to be done differently from the other course forms for the fall 2007 and spring 2008 semesters because the software could not run the analysis on the large amount of data and levels each variable contained. For these two semesters, averages were taken of g-theory analyses of 20 simple random samples with 2,500 records per sample. The results show that the variance estimates for the teacher universe are all higher than those for the course universe (see Table 4 and 5). The Standard Lecture Course Form is therefore more appropriate for making decisions on teachers than on courses for each semester.

Table 4

Standard Lecture Course Form s:c:t x *i G-theory Analysis*

	Spring 2007		Fall 2007		Spring 2008	
Source of Variation	Estimate	Percent	Estimate	Percent	Estimate	Percent
t	0.082	7.691%	0.124	12.533%	0.068	11.919%
i	0.012	1.106%	0.007	0.705%	0.005	0.809%
c:t	0.051	4.809%	0.058	5.879%	0.042	7.418%
t *x* i	0.013	1.188%	0.020	2.002%	0.021	3.600%
s:c:t	0.588	54.946%	0.412	41.796%	0.259	45.475%
c *x* i:t	0.011	1.030%	0.049	4.979%	0.029	5.042%
s:c:t *x* i, e	0.313	29.229%	0.316	32.105%	0.147	25.736%

Table 5

Standard Lecture Course Form s:t:c x *i G-theory Analysis*

	Spring 2007		Fall 2007		Spring 2008	
Source of Variation	Estimate	Percent	Estimate	Percent	Estimate	Percent
c	0.004	0.402%	0.005	0.496%	0.002	0.338%
i	0.012	1.105%	0.007	0.706%	0.005	0.824%
t:c	0.104	9.747%	0.159	16.114%	0.083	14.904%
c *x* i	0.008	0.728%	0.019	1.978%	0.013	2.280%
s:t:c	0.613	57.286%	0.424	43.042%	0.269	48.246%
t *x* i:c	0.013	1.183%	0.048	4.883%	0.027	4.828%
s:t:c *x* i, e	0.316	29.549%	0.323	32.781%	0.160	28.581%

The generalizability coefficients (g coefficients) show that the Standard Lecture Course Form's reliability is pretty stable over all three semesters (see Table 6). This reliability is excellent for most combinations of number of courses and students. Scenarios for small class sizes with a low number of students yield moderate reliability.

Table 6

Standard Lecture Course Form Estimated s:c:t x *i G Coefficients for 20 Items*

			Spring 2007		
			n_s		
n_c	5	10	15	20	25
1	0.322	0.421	0.470	0.498	0.517
2	0.486	0.592	0.638	0.664	0.680
3	0.586	0.684	0.724	0.746	0.760
4	0.653	0.741	0.776	0.795	0.807
5	0.701	0.781	0.812	0.828	0.838
6	0.737	0.810	0.837	0.852	0.861
7	0.765	0.831	0.856	0.869	0.877
			Fall 2007		
			n_s		
n_c	5	10	15	20	25
1	0.457	0.543	0.579	0.599	0.611
2	0.625	0.701	0.731	0.747	0.757
3	0.713	0.777	0.801	0.814	0.822
4	0.767	0.822	0.842	0.852	0.858
5	0.804	0.851	0.868	0.877	0.882
6	0.830	0.872	0.887	0.894	0.899
7	0.850	0.887	0.900	0.907	0.911
			Spring 2008		
			n_s		
n_c	5	10	15	20	25
1	0.409	0.488	0.521	0.539	0.551
2	0.578	0.652	0.681	0.697	0.707
3	0.671	0.735	0.759	0.772	0.780
4	0.729	0.785	0.806	0.816	0.823
5	0.769	0.818	0.836	0.845	0.851
6	0.798	0.842	0.858	0.866	0.871
7	0.820	0.860	0.874	0.881	0.885

A hierarchical linear model (HLM) was estimated to further understand how the variance components of the g-theory analysis can be disentangled into support for construct validity or support against construct validity. The random base model (see Table 7) for *Perceived Instructional Effectiveness* yielded reliability between sections of 0.443 and reliability across teachers of 0.630. These reliabilities were moderate and consistent with the results from the g-theory analysis. The estimates of *Perceived Instructional Effectiveness* between-section and within-section variability were 13.80887 and 118.17644 respectively, leading to an intraclass correlation (proportion of variance due to between-section differences) of 0.105 (see below).

Table 7

Standard Lecture Course Form Random Base Model: Random Effects

Random Effect	Variance	df	Chi-Square
	Level-1 and Level-2		
Between Section	13.809	628	1203.746
Within Section	118.176		
	Level-3		
Between Teachers	28.517	445	1442.372

Note. All random effects have a p-value < .001.

$$\rho = \frac{13.80887}{13.80887 + 118.17644} = 0.105$$

This intraclass correlation indicates that 10.5% of the variance in ratings is between sections, without any predictor variables. There is a significant amount of unexplained variance between teachers in the random base model, thus a full level-1 model should be estimated utilizing student characteristics (see Table 8). The percentage of between-section variance that is explained by the student measures in the level-1 model is 0.2% and the percentage of between-teacher variance that is explained by the student measures is 21.0% (see below).

Table 8

Standard Lecture Course Form Level-1 Model: Random Effects

Random Effect	Variance	df	Chi-Square
	Level-1 and Level-2		
Between Section	13.781	537	847.658
Within Section	93.463		
	Level-3		
Between Teachers	22.534	423	966.731

Note. All random effects have a p-value < .001.

$$\text{Proportion of between section variance} = \frac{(13.80887 - 13.78071)}{13.80887} = 0.002 \text{ or } 0.2\%$$

$$\text{Proportion of between teacher variance} = \frac{(28.51658 - 22.53355)}{28.51658} = 0.210 \text{ or } 21.0\%$$

A level-2 model was estimated (see Table 9) next utilizing course section characteristics since the between-teach variance was still significant in the level-1 model. The proportion of between-section variance that is explained by class section variables is 2.8% and the proportion of between-teacher variance that is explained by class section variables is 6.8% (see below).

Table 9

Standard Lecture Course Form Level-2 Model: Random Effects

Random Effect	Variance	df	Chi-Square
	Level-1 and Level-2		
Between Section	13.390	524	840.885
Within Section	92.612		
	Level-3		
Between Teachers	20.999	423	939.430

Note. All random effects have a p-value < .001.

$$\text{Proportion of between section variance} = \frac{(13.78071 - 13.38976)}{13.78071} = 0.028 \text{ or } 2.8\%$$

$$\text{Proportion of between teacher variance} = \frac{(22.53355 - 20.99877)}{22.53355} = 0.068 \text{ or } 6.8\%$$

The between-teacher variance is still significant, so the final step was to estimate the level-3 model utilizing teacher characteristics (see Table 10). The proportion of between-teacher variance explained by the additional teacher characteristics is 0.010 or 1.0% (see below). The percentage of between-teacher variance that at most can be explained by potential biases (sum of all percentages of models' between-teacher variances) is 28.8%.

Table 10

Standard Lecture Course Form Level-3 Model: Random Effects

Random Effect	Variance	df	Chi-Square
Level-1 and Level-2			
Between Section	13.307	524	842.322
Within Section	93.113		
Level-3			
Between Teachers	20.785	412	936.967

Note. All random effects have a p-value < .001.

$$\text{Proportion of between teacher variance} = \frac{(20.99877 - 20.78468)}{20.99877} = 0.010 \text{ or } 1.0\%$$

The Standard Lecture Course Form was the only course form that contained enough data for all three levels of the HLM model to be analyzed. Because of this, all potential biases’ hypotheses can be analyzed for this course form.

The variables that will be investigated for hypothesis 3 in order to determine the effect of students’ expected grades will be the student’s expected grade, overall GPA, actual class average grade, attempted hours, course difficulty relative to other courses, the interaction between interest level in the course and expected grade, the interaction between expected grade and overall GPA, the interaction between overall GPA and actual class average grade, and the interaction between expected grade and attempted hours. All expected grade hypotheses were investigated using the level-3 model for *Perceived Instructional Effectiveness*.

Expected grade did not have a significant association with ratings, giving no support for the “validity hypothesis.” This was also the same with the “student general motivation hypothesis,” where the interaction between expected grade and overall GPA (β = -0.729, t = -1.774, p = .076) and the interaction between expected grade and attempted hours (β = 0.063, t = 1.203, p = .230) were both not significant. The “students’ characteristics hypothesis” was also rejected because the interaction between expected grade and interest level (β = -0.246, t = -0.789, p = .430) was not significant. Course difficulty relative to other courses (β = 0.194, t = 0.495, p = .620) was not significant but interest level (β = -1.653, t = -8.348, p < .001) was, thereby giving some evidence supporting the “course specific motivation hypothesis.” The “grading-leniency hypothesis” was also supported because the interaction between overall GPA and actual class average grade was significant. Those with a higher overall GPA tended to give lower ratings and those in classes with a higher actual class average grade gave higher ratings. The “infer course quality hypothesis” could not be investigated because students’ actual grades were not in the data set.

Overall, only the “course specific motivation hypothesis” and “grading-leniency hypothesis” were supported. All other hypotheses, including the “validity hypothesis,” were rejected. These results point to students’ expected grades being a bias to ratings. It appears that one way teachers can get better ratings in their courses is by improving their students’ motivation for those courses. Also, addressing grade inflation can reduce bias and possibly reject the “grading-leniency hypothesis” in the future.

Hypothesis 4a was analyzed next. Teacher ethnicity, teacher age (β = -0.050, t = -1.540, p = .124), course level, course subject, course enrollment (β = -0.048, t = -1.572, p = .116), reason course was taken, course difficulty (β = 0.194, t = 0.495, p = .620), and attempted hours (β = 0.047, t = 1.559, p = .119) were all not significantly associated with *Perceived Instructional Effectiveness*. This is overwhelming support for hypothesis 4a. Unfortunately, there was a significant interaction between-teacher gender and student gender, as well as the student’s interest level in the course (β = -1.653, t = -8.348, p < .001). Male students tend to give lower ratings and male teachers tend to get higher ratings. These provide evidence against hypothesis 4a.

Most variables in models for *Rapport and Respect*, *Student Perceptions of Learning*, and *Perceived Instructional Effectiveness* failed to reject hypothesis 4b. The level of difficulty in the *Rapport and Respect* model (β = 0.037, t = 0.481, p = .630), *Student Perceptions of Learning* model (β = 0.051, t = 0.540, p = .589), and *Perceived Instructional Effectiveness* model (β = 0.194, t = 0.495, p = .620) were not significant. This implies that course difficulty does not positively or negatively impact ratings.

Next, hypothesis 4c was analyzed to determine if the amount of reading, amount of writing, and overall workload were associated with the *Rapport and Respect*, *Feedback and Accessibility*, and *Student Perceptions of Learning* factors. The hope is that those variables are positively associated with those factors. For *Rapport and Respect* and *Feedback and Accessibility*, the amount of reading (β = -0.025, t = -0.447, p = .654; β = -0.036, t = -0.594, p = .552) and amount of writing (β = 0.096, t = 1.767, p = .077; β = 0.031, t = 0.498, p = .618) were both not significant. However, overall workload writing (β = -0.190, t = -2.439, p = .015; β = -0.220, t = -2.363, p = .018) was significant for both of those factors. The *Student Perceptions of Learning* factor had a slightly different outcome. Both the amount of reading (β = -0.224, t = -0.797, p = .425) and the overall workload (β = -0.643, t = -1.487, p = .137) were not significant, but the amount of writing (β = 0.687, t = 2.372, p = .018) was. It appears that overall workload negatively impacts students' rapport and respect for their teacher as well as their perception of how accessible the teacher is and the quality of the teacher's feedback. The amount of writing negatively impacts students' perceptions of learning in the course.

There was also not enough evidence to reject hypothesis 4d since the pace of the course was not significant (β = 0.190, t = 1.214, p = .225) in the *Enthusiasm and Intellectual Stimulation* factor. Pace, therefore, did not have any impact on how enthusiastic and stimulated students were.

The number of hours per week spent outside of class was significantly associated with the *Rapport and Respect* and *Student Perceptions of Learning* factors. This is evident that the workload outside of class negatively influences students' respect for their teachers and their perceptions of how much they have learned. This is evidence against hypothesis 4e and implies that the workload outside of class is not fair and appropriate for classes requiring more than 11 hours of work outside of the classroom.

Finally, hypothesis 4f was analyzed in order to determine if the teachers with higher rank get better ratings or if rank is not associated to ratings. It is clear in the model for *Perceived Instructional Effectiveness* that the rank of the teacher does not impact ratings at all. This implies that teachers with more experience do not teach better or worse than teachers with less experience.

Seminar Course Form

The Seminar Course Form has good model fit based on the good NFI, CFI, and RMSEA values (see Table 11). The adequacy of this fit drops slightly, however, from the fall 2007 semester to the spring 2008 semester. It is likely that this is a result in the large drop in the number of responses between those two semesters. This drop did not impact model fit much, but increasing the number of responses through increasing the response rate will ensure that construct validity remains good.

Table 11
Seminar Course Form Goodness-of-fit Indices

Fall 2007					
Model	N	NPAR	NFI	CFI	RMSEA
One-Factor	1585	21	0.888	0.894	0.105
Five-Factor Orthogonal	1585	25	0.663	0.667	0.186
Five-Factor Oblique	1585	25	0.940	0.945	0.078
Second-Order	1585	26	0.935	0.940	0.081
Spring 2008					
Model	N	NPAR	NFI	CFI	RMSEA
One-Factor	216	21	0.818	0.848	0.136
Five-Factor Orthogonal	216	25	0.632	0.654	0.206
Five-Factor Oblique	216	25	0.875	0.905	0.111
Second-Order	216	26	0.870	0.901	0.114

Note. All chi-square values are significant at α = .001.

All questions loaded positively and significantly at the alpha = .05 level in their expected first-order factors (see Table 12). The same followed for all first-order factor loadings (see Table 13). This means that every question contributes to the construct validity of this course form. Despite this, changing questions 2 and 7, which have lower structure coefficients than other questions within their factors, might improve construct validity. These changes are not absolutely necessary though.

Table 12
Seminar Course Form Pattern Structure Coefficients

		Fall 2007		Spring 2008	
Factor	Question	Pattern Coefficient	Structure Coefficient	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1	0.450	0.787	0.326	0.804
	2	0.488	0.782	0.331	0.743
	3	0.512	0.812	0.449	0.864
	4	0.615	0.855	0.533	0.909
Enthusiasm and Intellectual Stimulation	5	0.563	0.880	0.457	0.868
	6	0.550	0.824	0.491	0.858
	7	0.390	0.756	0.279	0.724
	8	0.529	0.864	0.397	0.830
Rapport and Respect	9	0.536	0.818	0.457	0.832
	10	0.499	0.619	0.461	0.847
	11	0.455	0.862	0.346	0.863
	12	0.476	0.817	0.390	0.820
Feedback and Accessibility	13	0.397	0.745	0.383	0.804
	14	0.510	0.860	0.336	0.854
	15	0.528	0.826	0.442	0.853
	16	0.494	0.844	0.426	0.903
Student Perceptions of Learning	17	0.497	0.871	0.449	0.893
	18	0.530	0.808	0.421	0.847
	19	0.531	0.903	0.423	0.909
	20	0.481	0.863	0.384	0.891

Note. All coefficients are significant at the α = .05 level.

Table 13
Seminar Course Form Pattern and Structure Coefficients
for Perceived Instructional Effectiveness

	Fall 2007		Spring 2008	
First-Order Factor	Pattern Coefficient	Structure Coefficient	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1.191	0.937	1.201	0.945
Enthusiasm and Intellectual Stimulation	1.393	0.971	1.338	0.979
Rapport and Respect	1.212	0.915	1.107	0.897
Feedback and Accessibility	1.282	0.919	1.206	0.908
Student Perceptions of Learning	1.350	0.951	1.311	0.967

Note. All coefficients are significant at the α = .05 level.

The g-theory analysis clearly shows that the Seminar Course Form is more reliable for making decisions about teachers instead of courses (see Tables 14 and 15). The g coefficients for different scenarios based on the number of courses a teacher teaches and the number of student in each class drop slightly from the fall 2007 semester to the spring 2008 semester, also possibly due to the decreased sample size (see Table 16). Despite this, the coefficients show moderate to good reliability.

Table 14
Seminar Course Form s:c:t x *i G-theory Analysis*

	Fall 2007		Spring 2008	
Source of Variation	Estimate	Percent	Estimate	Percent
t	0.091	12.302%	0.082	19.231%
i	0.016	2.212%	0.006	1.367%
c:t	0.007	1.012%	0.050	11.824%
t *x* i	0.013	1.754%	0.004	1.048%
s:c:t	0.328	44.345%	0.141	33.285%
c *x* i:t	0.006	0.772%	0.010	2.318%
s:c:t *x* i, e	0.278	37.603%	0.131	30.928%

Table 15
Seminar Course Form s:t:c x *i G-theory Analysis*

	Fall 2007		Spring 2008	
Source of Variation	Estimate	Percent	Estimate	Percent
c	0.000	0.054%	0.017	3.921%
i	0.015	2.089%	0.006	1.381%
t:c	0.091	12.257%	0.093	22.026%
c *x* i	0.010	1.352%	0.007	1.591%
s:t:c	0.335	45.300%	0.158	37.549%
t *x* i:c	0.009	1.206%	0.004	0.941%
s:t:c *x* i, e	0.279	37.741%	0.137	32.591%

Table 16

Seminar Course Form Estimated s:c:t x *i G Coefficients for 20 Items*

	n_s									
	Fall 2007					Spring 2008				
n_c	5	10	15	20	25	5	10	15	20	25
1	0.542	0.681	0.745	0.781	0.805	0.504	0.554	0.573	0.583	0.590
2	0.701	0.808	0.851	0.874	0.889	0.669	0.712	0.728	0.736	0.741
3	0.778	0.861	0.894	0.911	0.921	0.752	0.787	0.800	0.807	0.810
4	0.822	0.891	0.917	0.930	0.938	0.801	0.831	0.842	0.847	0.850
5	0.851	0.910	0.931	0.942	0.949	0.834	0.860	0.869	0.873	0.876
6	0.872	0.923	0.941	0.950	0.956	0.857	0.880	0.888	0.892	0.894
7	0.888	0.932	0.948	0.956	0.961	0.875	0.895	0.902	0.906	0.908

An HLM model was estimated to further understand how the variance components of the g-theory analysis can be disentangled into support for construct validity or support against construct validity. The random base model (see Table 17) for *Perceived Instructional Effectiveness* yielded a reliability between sections of 0.389 and a reliability across teachers of 0.651, which are moderate but lower than the g-theory results. The estimates of the variability of *Perceived Instructional Effectiveness* between-section and within-sections were 7.466 and 80.134 respectively leading to an intraclass correlation (proportion of variance due to between-section differences) of 0.085 (see below).

Table 17

Seminar Course Form Random Base Model: Random Effects

Random Effect	Variance	df	Chi-Square
Level-1 and Level-2			
Between Section	7.466	104	179.518
Within Section	80.134		
Level-3			
Between Teachers	22.359	79	287.341

Note. All random effects have a p-value < .001.

$$\rho = \frac{7.46587}{7.46587 + 80.13416} = 0.085$$

This intraclass correlation indicates that 8.5% of the variance in ratings is between sections, without any predictor variables. The between-teacher variance was statistically significant, meaning that higher-level models can explain more differences. The next model estimated was the level-1 model utilizing student characteristics (see Table 18). The between-section variance that is explained by the student measures in the level-1 model is 0.075 or 7.5%.

Table 18

Seminar Course Form Level-1 Model: Random Effects

Random Effect	Variance	df	Chi-Square	P-Value
	Level-1 and Level-2			
Between Section	1.315	102	66.654	.104
Within Section	74.135			
	Level-3			
Between Teachers	23.806	78	253.239	<.001

$$\text{Proportion of explained variance} = \frac{(80.13416 - 74.13454)}{80.13416} = 0.075 \text{ or } 7.5\%$$

The significant between-teacher variance in the level-1 model suggests that there is more that can be explained by a level-2 model, but there is not enough data to estimate a level-2 model. It appears that potential biases only explain a small percentage of *Perceived Instructional Effectiveness*. The level-1 model can be used to evaluate some potential biases hypotheses.

Hypothesis 4a was analyzed first. The pace of the course (β = 0.340, t = 0.301, p = .764), attempted hours (β = -0.051, t = -1.354, p = .176), and the reason the course was taken were all not significantly associated with *Perceived Instructional Effectiveness*. The level of interest (β = -1.226, t = -3.845, p < .001) was significantly associated with *Perceived Instructional Effectiveness*. It appears that most of the evidence fails to reject hypothesis 4a, but those who have a higher interest level in the course will tend to give lower ratings. This might be due to these students having higher expectations for the course and the teacher.

The variables in the level-1 models for *Rapport and Respect*, *Student Perceptions of Learning*, and *Perceived Instructional Effectiveness* failed to reject hypothesis 4b. The level of difficulty in the *Rapport and Respect* model (β = -0.156, t = -1.169, p = .243), *Student Perceptions of Learning* model (β = 0.021, t = 0.123, p = .903), and *Perceived Instructional Effectiveness* model (β = -0.521, t = -0.803, p = .422) were all not significant. This implies that course difficulty does not negatively impact ratings, but unfortunately also implies that courses with higher difficulty do not promote learning and students do not have positive or negative respect for teachers who challenge them.

The amount of reading, amount of writing, and overall workload were analyzed next to determine if they were associated with the factors of *Rapport and Respect*, *Feedback and Accessibility*, and *Student Perceptions of Learning*. If these variables are not significant then there is no evidence for or against hypothesis 4c, but if there is a positive association then these variables provide evidence for hypothesis 4c. For the factor of *Rapport and Respect* the amount of reading (β = 0.016, t = 0.192, p = .848), amount of writing (β = -0.002, t = -0.015, p = .988), and overall workload (β = 0.00009, t = 0.001, p = .999) were not significant. The factor of *Feedback and Accessibility* had a similar outcome: the amount of reading (β = 0.019, t = 0.211, p = .833), amount of writing (β = 0.031, t = 0.289, p = .772), and overall workload (β = -0.149, t = -1.293, p = .197) were not significant. These two factors show no evidence for or against hypothesis 4c. The amount of reading (β = 0.063, t = 0.606, p = .544) and amount of writing (β = 0.089, t = 0.763, p = .446) were also not significant for the factor of *Student Perceptions of Learning*, but unfortunately overall workload (β = -0.276, t = -2.204, p = .028) was significant. For the most part, course workload does not influence students' respect for their teachers and their teachers' ability to give feedback but does negatively influence their perception of how much they learned in the course.

There was also not enough evidence for or against hypothesis 4d since the pace of the course was not significant (β = 0.157, t = 0.596, p = .551) in the *Enthusiasm and Intellectual*

Stimulation factor. Also, the number of hours per week spent outside of class was not significantly associated with the *Rapport and Respect* and *Student Perceptions of Learning* factors. This is evidence that the workload outside of class does not influence students' respect for their teachers and their perceptions of what they learned in the course, giving no support for or against hypothesis 4e.

Online Course Form

The data for the Online Course Form had good model fit for each semester (see Table 19). This construct validity was stable over each semester, which is good considering the fluctuation in the number of observations. Each question loaded positively and significantly on its expected first-order factor (see Table 20), and each first-order factor loaded positively and significantly on *Perceived Instructional Effectiveness* (see Table 21). This is further evidence for construct validity and it appears that no questions need to be changed to improve model fit.

Table 19

Online Course Form Goodness-of-fit Indices

Spring 2007					
Model	N	NPAR	NFI	CFI	RMSEA
One-Factor	173	21	0.865	0.892	0.141
Five-Factor Orthogonal	173	25	0.664	0.684	0.241
Five-Factor Oblique	173	25	0.900	0.927	0.120
Second-Order	173	26	0.897	0.923	0.122
Fall 2007					
Model	N	NPAR	NFI	CFI	RMSEA
One-Factor	1171	21	0.879	0.883	0.139
Five-Factor Orthogonal	1171	25	0.714	0.718	0.216
Five-Factor Oblique	1171	25	0.945	0.950	0.094
Second-Order	1171	26	0.943	0.948	0.096
Spring 2008					
Model	N	NPAR	NFI	CFI	RMSEA
One-Factor	796	21	0.866	0.874	0.124
Five-Factor Orthogonal	796	25	0.678	0.685	0.197
Five-Factor Oblique	796	25	0.923	0.932	0.094
Second-Order	796	26	0.922	0.930	0.096

Note. All chi-square values are significant at α = .001.

Table 20

Online Course Form Pattern and Structure Coefficients

Spring 2007			
Factor	Question	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1	0.673	0.903
	2	0.614	0.825
	3	0.655	0.899
	4	0.627	0.838
Enthusiasm and Intellectual Stimulation	5	0.570	0.953
	6	0.577	0.929
	7	0.542	0.857
	8	0.610	0.946
Rapport and Respect	9	0.617	0.956
	10	0.613	0.958
	11	0.584	0.894
	12	0.614	0.908
Feedback and Accessibility	13	0.665	0.914
	14	0.621	0.892
	15	0.586	0.897
	16	0.646	0.909
Student Perceptions of Learning	17	0.532	0.919
	18	0.569	0.914
	19	0.577	0.935
	20	0.559	0.962

Fall 2007			
Factor	Question	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1	0.620	0.863
	2	0.588	0.865
	3	0.604	0.910
	4	0.589	0.895
Enthusiasm and Intellectual Stimulation	5	0.592	0.949
	6	0.609	0.960
	7	0.557	0.880
	8	0.612	0.936
Rapport and Respect	9	0.592	0.934
	10	0.576	0.926
	11	0.584	0.911
	12	0.600	0.911
Feedback and Accessibility	13	0.596	0.797
	14	0.570	0.831
	15	0.585	0.916
	16	0.613	0.906
Student Perceptions of Learning	17	0.602	0.932
	18	0.584	0.895
	19	0.601	0.895
	20	0.623	0.948

Table 20 (Continued)

Spring 2008			
Factor	Question	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1	0.439	0.803
	2	0.433	0.781
	3	0.414	0.853
	4	0.360	0.758
Enthusiasm and Intellectual Stimulation	5	0.430	0.916
	6	0.447	0.915
	7	0.343	0.805
	8	0.482	0.896
Rapport and Respect	9	0.365	0.871
	10	0.355	0.852
	11	0.376	0.822
	12	0.444	0.863
Feedback and Accessibility	13	0.509	0.772
	14	0.443	0.798
	15	0.453	0.903
	16	0.492	0.911
Student Perceptions of Learning	17	0.424	0.908
	18	0.354	0.795
	19	0.430	0.866
	20	0.433	0.917

Note. All coefficients are significant at the α = .05 level.

Table 21

Online Course Form Pattern and Structure Coefficients for Perceived Instructional Effectiveness

Spring 2007		
First-Order Factor	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1.510	0.984
Enthusiasm and Intellectual Stimulation	1.832	0.960
Rapport and Respect	1.647	0.962
Feedback and Accessibility	1.661	0.972
Student Perceptions of Learning	1.881	0.979
Fall 2007		
First-Order Factor	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1.429	0.918
Enthusiasm and Intellectual Stimulation	1.635	0.955
Rapport and Respect	1.564	0.967
Feedback and Accessibility	1.477	0.929
Student Perceptions of Learning	1.555	0.982
Spring 2008		
First-Order Factor	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1.176	0.930
Enthusiasm and Intellectual Stimulation	1.375	0.961
Rapport and Respect	1.261	0.933
Feedback and Accessibility	1.187	0.887
Student Perceptions of Learning	1.389	0.976

Note. All coefficients are significant at the $\alpha = .05$ level.

The g-theory analysis shows evidence that the Online Course Form is more reliable in making decisions about teachers than for courses (see Table 22 and 23). The g coefficients were their highest, showing good reliability, in the spring 2007 semester and then dropped to poor to moderate reliability (see Table 24). Changes in the structure of online courses and technology should be looked at, if there are any, to see if reliability can be increased.

Table 22

Online Course Form s:c:t x *i G-theory Analysis*

	Spring 2007		Fall 2007		Spring 2008	
Source of Variation	Estimate	Percent	Estimate	Percent	Estimate	Percent
t	0.108	8.610%	0.101	7.964%	0.049	10.214%
i	0.009	0.725%	0.010	0.783%	0.006	1.169%
c:t	0.000	0.000%	0.739	58.372%	0.097	20.325%
t *x* i	0.026	2.088%	0.026	2.040%	0.009	1.826%
s:c:t	0.807	64.211%	0.053	4.221%	0.160	33.452%
c *x* i:t	0.000	0.000%	0.000	0.000%	0.004	0.881%
s:c:t *x* i, e	0.306	24.366%	0.337	26.618%	0.153	32.132%

Table 23

Online Course Form s:t:c x *i G-theory Analysis*

	Spring 2007		Fall 2007		Spring 2008	
Source of Variation	Estimate	Percent	Estimate	Percent	Estimate	Percent
c	0.021	1.695%	0.000	0.000%	0.037	7.692%
i	0.010	0.791%	0.010	0.788%	0.006	1.174%
t:c	0.067	5.455%	0.817	64.431%	0.085	17.748%
c *x* i	0.030	2.458%	0.006	0.456%	0.000	0.000%
s:t:c	0.795	65.010%	0.084	6.663%	0.181	37.525%
t *x* i:c	0.000	0.000%	0.008	0.634%	0.015	3.098%
s:t:c *x* i, e	0.301	24.591%	0.343	27.027%	0.158	32.764%

Table 24

Online Course Form Estimated s:c:t x *i G Coefficients for 20 Items*

	Spring 2007				
	n_s				
n_c	5	10	15	20	25
1	0.395	0.564	0.658	0.718	0.760
2	0.564	0.718	0.790	0.832	0.859
3	0.658	0.790	0.847	0.878	0.898
4	0.718	0.832	0.878	0.903	0.919
5	0.760	0.859	0.898	0.919	0.932
6	0.790	0.878	0.912	0.930	0.941
7	0.814	0.892	0.922	0.938	0.947
	Fall 2007				
	n_s				
n_c	5	10	15	20	25
1	0.118	0.119	0.119	0.119	0.119
2	0.211	0.212	0.213	0.213	0.213
3	0.286	0.287	0.288	0.288	0.289
4	0.347	0.349	0.350	0.350	0.351
5	0.399	0.401	0.402	0.402	0.403
6	0.443	0.445	0.446	0.446	0.447
7	0.481	0.483	0.484	0.484	0.485
	Spring 2008				
	n_s				
n_c	5	10	15	20	25
1	0.271	0.299	0.309	0.315	0.318
2	0.426	0.459	0.472	0.478	0.482
3	0.526	0.559	0.571	0.578	0.582
4	0.596	0.628	0.639	0.645	0.649
5	0.647	0.677	0.688	0.693	0.697
6	0.687	0.715	0.725	0.730	0.733
7	0.718	0.745	0.754	0.759	0.761

An HLM model was estimated to further understand how the variance components of the g-theory analysis can be disentangled into support for construct validity or support against construct validity. The random base model (see Table 25) for *Perceived Instructional Effectiveness* yielded a reliability between sections of 0.002 and a reliability across teachers of

0.582. These reliabilities were poor to moderate at best. These results are pretty consistent with the g-theory analysis. The estimates of between-section and within-section variability of *Perceived Instructional Effectiveness* were 0.050 and 194.621 respectively, leading to an intraclass correlation (proportion of variance due to between-section differences) of 0.0003 (see below).

Table 25

Online Course Form Random Base Model: Random Effects

Random Effect	Variance	df	Chi-Square	P-Value
	Level-1 and Level-2			
Between Section	0.050	69	72.130	.375
Within Section	194.621			
	Level-3			
Between Teachers	41.881	89	493.022	<.001

$$\rho = \frac{0.04978}{0.04978 + 194.62057} = 0.0003$$

This intraclass correlation indicates that 0.03% of the variance in ratings is between sections, without any predictor variables. The between-teacher variance was statistically significant, so higher-level models can explain more differences between teachers. The next model estimated was the level-1 model utilizing student characteristics (see Table 26). The between-section variance that is explained by the student measures in the level-1 model is 9.6% and the between-teacher variance explained by the student measures is 9.7% (see below).It seems that again, potential biases only explain a small percentage of *Perceived Instructional Effectiveness*.

Table 26

Online Course Form Level-1 Model: Random Effects

Random Effect	Variance	df	Chi-Square	P-Value
	Level-1 and Level-2			
Between Section	1.549	62	66.654	.320
Within Section	175.984			
	Level-3			
Between Teachers	37.800	85	253.239	<.001

$$\text{Proportion of between section variance} = \frac{(194.62057 - 175.98352)}{194.62057} = 0.096 \text{ or } 9.6\%$$

$$\text{Proportion of between teacher variance} = \frac{(41.88136 - 37.79965)}{41.88136} = 0.097 \text{ or } 9.7\%$$

The significant between-teacher variance suggests that there is more that can be explained by a level-2 model, but unfortunately there is not enough data for a level-2 model. However, the level-1 model can be used to evaluate some potential biases hypotheses.

Hypothesis 4a was analyzed first. It is apparent that this hypothesis should be rejected from the level-1 model for *Perceived Instructional Effectiveness*. The reason for taking the course, interest level (β = -2.397, t = -3.368, p = .001), and attempted hours (β = -0.330, t = -3.954, p < .001) were all negatively associated with *Perceived Instructional Effectiveness*. Those with more interest in the course might be giving lower ratings due to having higher expectations. There was some support for hypothesis 4a in that the pace of the course did not influence ratings (β = -1.721, t = -1.331, p = .184).

Most variables in the level-1 models for *Rapport and Respect*, *Student Perceptions of Learning*, and *Perceived Instructional Effectiveness* failed to reject hypothesis 4b. The level of difficulty in the *Rapport and Respect* model (β = -0.005, t = -0.026, p = .980), *Student Perceptions of Learning* model (β = 0.135, t = 0.683, p = .494), and *Perceived Instructional Effectiveness* model (β = -0.208, t = -0.224, p = .823) were not significant. This implies that course difficulty does not negatively impact ratings, but unfortunately also implies that courses with higher difficulty do not promote learning and students do not have positive or negative respect for teachers who challenge them. The level of difficulty, therefore, does not provide evidence for or against construct validity.

Next, hypothesis 4c was analyzed to determine if the amount of reading, amount of writing, and overall workload were associated with the *Rapport and Respect*, *Feedback and Accessibility*, and *Student Perceptions of Learning* factors. The hope is that those variables are either not associated or positively associated with those factors. The amount of reading (β = 0.023, t = 0.113, p = .911), amount of writing (β = -0.003, t = -0.015, p = .988), and overall workload (β = -0.160, t = -0.639, p = .523) were not significant with the *Rapport and Respect* factor. The amount of reading (β = -0.023, t = -0.122, p = .903), amount of writing (β = -0.119, t = -0.754, p = .451), and overall workload (β = -0.249, t = -1.181, p = .238) were also not significant with the *Feedback and Accessibility* factor. The amount of reading (β = -0.193, t = -0.889, p = .375), amount of writing (β = -0.052, t = -0.370, p = .711), and overall workload (β = -0.042, t = -0.639, p = .178) were not significant with the *Student Perceptions of Learning* factor. These variables offer no evidence for or against construct validity.

There was also not enough evidence to reject hypothesis 4d since the pace of the course was not significant (β = -0.214, t = -0.914, p = .361) in the *Enthusiasm and Intellectual Stimulation* factor. The pace of the course then is not a bias, but unfortunately does not provide evidence that faster courses offer more enthusiasm and stimulation.

The number of hours per week spent outside of class was not significantly associated with the *Rapport and Respect* factor but was significantly and negatively associated with the *Student Perceptions of Learning* factor. This is evidence that the workload outside of class does not influence students' respect for their teachers but does negatively influence their perceptions of how much they have learned.

Activities Course Form

Model fit for the Activities Course Form was good for the fall 2007 and spring 2008 semesters, as can be seen in Table 27. There was better model fit in fall 2007 though. It is difficult to determine why the fit was not as good for the spring 2008 semester due to the vast amount of changes that occurred. There were more policies and procedures in place for the spring 2008 semester, the number of response options changed from five to four, and the number of observations dropped. Increasing the response rate, which should increase the sample size, would be a good first step in improving construct validity. It should also be noted that there is still much room for improvement on the Activities Course Form since the NFI and CFI fall short of their ideal cutoff values.

Table 27
Activities Course Form Goodness-of-fit Indices

Fall 2007					
Model	N	NPAR	NFI	CFI	RMSEA
One-Factor	406	21	0.794	0.809	0.156
Five-Factor Orthogonal	406	25	0.671	0.683	0.201
Five-Factor Oblique	406	25	0.890	0.906	0.113
Second-Order	406	26	0.884	0.900	0.116
Spring 2008					
Model	N	NPAR	NFI	CFI	RMSEA
One-Factor	119	21	0.766	0.796	0.205
Five-Factor Orthogonal	119	25	0.600	0.622	0.279
Five-Factor Oblique	119	25	0.825	0.856	0.177
Second-Order	119	26	0.818	0.848	0.182

Note. All chi-square values are significant at α = .001.

The pattern and structure coefficients of each question can be seen in Table 28. These define how the 20 items can be linearly combined to form each of the five proposed factors. Each item is significant at an alpha level of .05 in this model for both the fall 2007 and spring 2008 semester, which is also evidence of construct validity. Unfortunately, question 8 is negative in the fall 2007 semester, which is evidence against construct validity. Given that the loading for this question changed from negative to positive in the spring 2008 semester, it seems likely that the negative coefficient was due to an unknown reliability or construct validity problem in the fall 2007 semester. It might be beneficial regardless to change this question. Upon comparing the structure coefficients within each first-order factor, changing questions 2, 7, 12, and 13 might also improve construct validity.

Table 28

Activities Course Form Pattern and Structure Coefficients

		Fall 2007		Spring 2008	
Factor	Question	Pattern Coefficient	Structure Coefficient	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1	0.379	0.776	0.551	0.939
	2	0.622	0.905	0.435	0.835
	3	0.617	0.924	0.575	0.966
	4	0.576	0.797	0.447	0.857
Enthusiasm and Intellectual Stimulation	5	0.436	0.761	0.384	0.927
	6	0.577	0.864	0.390	0.942
	7	0.516	0.874	0.327	0.828
	8	-0.463	-0.804	0.376	0.937
Rapport and Respect	9	0.558	0.896	0.504	0.966
	10	0.542	0.811	0.499	0.915
	11	0.497	0.887	0.437	0.888
	12	0.403	0.545	0.347	0.694
Feedback and Accessibility	13	0.500	0.847	0.369	0.746
	14	0.530	0.879	0.416	0.796
	15	0.555	0.858	0.388	0.799
	16	0.565	0.899	0.635	0.959
Student Perceptions of Learning	17	0.498	0.866	0.480	0.988
	18	0.493	0.907	0.479	0.986
	19	0.558	0.947	0.444	0.943
	20	0.555	0.921	0.437	0.957

Note. All coefficients are significant at the α = .05 level.

Similarly, the pattern and structure coefficients of the first-order factors were computed (see Table 29). Each first-order factor loading was positive and significant at the alpha = .05 level. The *Enthusiasm and Intellectual Stimulation* factor seems to play a large role in *Perceived Instructional Effectiveness* for the spring 2008 semester, indicating that teachers with more enthusiasm and more interesting activities tend to be better teachers for activities courses.

Table 29

Activities Course Form Pattern and Structure Coefficients for Perceived Instructional Effectiveness

	Fall 2007		Spring 2008	
First-Order Factor	Pattern Coefficient	Structure Coefficient	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1.096	0.878	1.173	0.938
Enthusiasm and Intellectual Stimulation	1.288	0.965	1.827	1.000
Rapport and Respect	1.378	0.948	1.452	0.958
Feedback and Accessibility	1.327	0.934	1.154	0.980
Student Perceptions of Learning	1.249	0.903	1.540	0.973

Note. All coefficients are significant at the α = .05 level.

The next step in the analysis was to move from analyzing the construct validity to analyzing the reliability of the Activities Course Form as it pertains to making decisions about the teacher and about the course. Table 30 contains the variance estimates and percentages for the g-theory model measuring the reliability to make decisions on the teacher (i.e., the $s : c : t \times i$ model) and Table 31 contains the variance estimates and percentages for the model measuring the reliability to make decisions on the course (i.e., the $s : t : c \times i$ model). It is clear from these data that the faculty variance is higher than the course variance for the fall 2007 semester and the Activities Course Form is therefore more reliable for making decisions about teachers for this semester. The reverse is true for spring 2008, where the Activities Course Form is more reliable for making decisions about courses. This is a reliability problem.

Table 30

Activities Course Form G-theory s:c:t x *i Analysis*

	Fall 2007		Spring 2008	
Source of Variation	Estimate	Percent	Estimate	Percent
t	0.142	10.482%	0.022	3.896%
i	0.363	26.823%	0.001	0.223%
c:t	0.000	0.000%	0.000	0.000%
t *x* i	0.000	0.000%	0.005	0.889%
s:c:t	0.303	22.422%	0.409	72.287%
c *x* i:t	0.191	14.100%	0.000	0.000%
s:c:t *x* i, e	0.354	26.173%	0.128	22.705%

Table 31
Activities Course Form G-theory s:t:c x *i Analysis*

	Fall 2007		Spring 2008	
Source of Variation	Estimate	Percent	Estimate	Percent
c	0.000	0.000%	0.073	11.672%
i	0.357	26.479%	0.001	0.193%
t:c	0.134	9.910%	0.000	0.000%
c *x* i	0.145	10.779%	0.005	0.797%
s:t:c	0.277	20.526%	0.415	66.713%
t *x* i:c	0.000	0.000%	0.000	0.000%
s:t:c *x* i, e	0.435	32.306%	0.128	20.625%

A decision study for the model based on the teacher was also part of the g-theory analysis. This decision study yields g coefficients for separate scenarios of the number of courses a teacher might teach and the number of students that might be in those courses for an instrument consisting of 20 items (see Table 32). The reliability of the Activities Course Form increased from the fall 2007 semester to the spring 2008 semester despite the drop in construct validity. Overall, reliability of this form was good.

Table 32
Activities Course Form Estimated s:c:t x *i G Coefficients for 20 Items*

	n_s									
	Fall 2007					Spring 2008				
n_c	5	10	15	20	25	5	10	15	20	25
1	0.658	0.773	0.821	0.847	0.864	0.209	0.345	0.441	0.512	0.566
2	0.794	0.872	0.902	0.917	0.927	0.345	0.512	0.610	0.674	0.720
3	0.852	0.911	0.932	0.943	0.950	0.441	0.610	0.699	0.754	0.792
4	0.885	0.932	0.948	0.957	0.962	0.512	0.674	0.754	0.802	0.833
5	0.906	0.945	0.958	0.965	0.969	0.566	0.720	0.792	0.833	0.860
6	0.920	0.953	0.965	0.971	0.974	0.610	0.754	0.819	0.856	0.879
7	0.931	0.960	0.970	0.975	0.978	0.645	0.781	0.840	0.873	0.894

An HLM model was estimated next to further understand how the variance components of the g-theory analysis can be disentangled into support for construct validity or support against construct validity. The random base model (see Table 33) for *Perceived Instructional Effectiveness* yielded reliability between sections of 0.152 and a reliability across teachers of 0.470. These reliabilities were much poorer than the g-theory reliabilities. The estimates of between-section and within-section variability of *Perceived Instructional Effectiveness* were 2.743 and 80.306 respectively, leading to an intraclass correlation (proportion of variance due to between-section differences) of 0.033 (see below).

Table 33

Activities Course Form Random Base Model: Random Effects

Random Effect	Variance	df	Chi-Square	P-Value
	Level-1 and Level-2			
Between Section	2.743	29	31.330	.350
Within Section	80.306			
	Level-3			
Between Teachers	11.058	35	80.197	<.001

$$\rho = \frac{2.74306}{2.74306 + 80.30559} = 0.033$$

This intraclass correlation is the proportion of variance in ratings that is accounted for between sections, without any predictor variables being specified. The proportion of variance explained here is 3.3%. The between-teacher variance was statistically significant, indicating that higher-level models can explain more differences in this case. Unfortunately there are not enough data to estimate any higher-level models. It would appear that models with potential biases will only explain a very small amount of ratings.

Independent Research Course Form

There were only enough data to run the analysis on the Independent Research Course Form for the spring 2008 semester, and even for this semester the number of observations was low. As can be seen in Table 34, this form does not show much evidence of construct validity. The NFI, CFI, and RMSEA are far away from the ideal cutoff values. Given the low number of observations, increasing the response rate will be necessary in order to increase validity.

Table 34
Independent Research Course Form Goodness-of-fit Indices

Spring 2008					
Model	N	NPAR	NFI	CFI	RMSEA
One-Factor	89	21	0.615	0.665	0.207
Five-Factor Orthogonal	89	25	0.533	0.576	0.233
Five-Factor Oblique	89	25	0.731	0.788	0.170
Second-Order	89	26	0.723	0.779	0.173

Note. All chi-square values are significant at α = .001.

Each question significantly loaded on its proposed first-order factor at the alpha = .05 level and was positive (see Table 35). Upon comparing the structure coefficients for questions within each factor, questions 1, 2, 5, 8, 9, 10, and 13 might be good places for improvement. Each first-order factor loading on the *Perceived Instructional Effectiveness* factor was also positive and significant (see Table 36). The *Enthusiasm and Intellectual Stimulation* factor seems to play a large role in *Perceived Instructional Effectiveness,* showing that teachers with more enthusiasm and who participate in their students' projects are better independent research teachers.

Table 35
Independent Research Course Form Pattern and Structure Coefficients

Spring 2008			
Factor	Question	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1	0.301	0.781
	2	0.319	0.800
	3	0.480	0.921
	4	0.380	0.890
Enthusiasm and Intellectual Stimulation	5	0.206	0.667
	6	0.337	0.832
	7	0.302	0.767
	8	0.328	0.520
Rapport and Respect	9	0.195	0.805
	10	0.203	0.769
	11	0.253	0.958
	12	0.291	0.976
Feedback and Accessibility	13	0.313	0.750
	14	0.471	0.957
	15	0.289	0.830
	16	0.386	0.833
Student Perceptions of Learning	17	0.371	0.778
	18	0.335	0.842
	19	0.415	0.763
	20	0.321	0.845

Note. All coefficients are significant at the α = .05 level.

Table 36
Independent Research Course Form Pattern and Structure Coefficients for Perceived Instructional Effectiveness

Spring 2008		
First-Order Factor	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1.149	0.905
Enthusiasm and Intellectual Stimulation	1.152	1.000
Rapport and Respect	1.153	0.812
Feedback and Accessibility	1.092	0.922
Student Perceptions of Learning	1.185	0.944

Note. All coefficients are significant at the α = .05 level.

The reliability of the Independent Research Course Form is also problematic. The g-theory analysis shows that this form explains virtually no variability amongst teachers (see Table 37). The analysis also shows that this form explains more variability amongst courses than amongst teachers (see Table 38). This means that students are answering the questions more based on their perception of the course than the teacher, which is evidence against construct validity. This might be due to the increased amount of independent learning without teacher input in independent research courses. The g coefficients for different scenarios of number of courses and students were all 0.000, illustrating that this form is highly unreliable and should be modified (see Table 39). HLM models were not created due to the low number of observations on this course form.

Table 37

Independent Research Course Form Spring 2008 s:c:t x *i G-theory Analysis*

Source of Variation	Estimate	Percent
t	0.003	0.947%
i	0.007	1.961%
c:t	0.201	60.287%
t *x* i	0.003	0.963%
s:c:t	0.000	0.000%
c *x* i:t	0.014	4.183%
s:c:t *x* i, e	0.106	31.659%

Table 38

Independent Research Course Form Spring 2008 s:t:c x *i G-theory Analysis*

Source of Variation	Estimate	Percent
c	0.006	2.394%
i	0.006	2.370%
t:c	0.122	45.195%
c *x* i	0.002	0.723%
s:t:c	0.000	0.000%
t *x* i:c	0.009	3.308%
s:t:c *x* i, e	0.124	46.011%

Table 39
Independent Research Course Form Spring 2008 Estimated s:c:t x *i G Coefficients for 20 Items*

	n_s				
n_c	5	10	15	20	25
1	0.015	0.015	0.015	0.015	0.015
2	0.030	0.030	0.030	0.030	0.030
3	0.045	0.045	0.045	0.045	0.045
4	0.058	0.059	0.059	0.059	0.059
5	0.072	0.072	0.072	0.072	0.072
6	0.085	0.085	0.085	0.085	0.085
7	0.098	0.098	0.098	0.098	0.098

Internship, Practica, and Clinical Course Form

The Internship, Practica, and Clinical Course Form has good construct validity as can be seen by the NFI, CFI, and RMSEA values, which are close to their ideal cutoff values (see Table 40). This course form has enough data to analyze both the fall 2007 and spring 2008, and it is clear that the construct validity is consistent between these two semesters. Each question's factor loading (see Table 41) and each first-order factor's loading were positive and significant (see Table 42), providing further evidence for construct validity. It appears that this course form does not need to be changed.

Table 40

Internship, Practica, and Clinical Course Form Goodness-of-fit Indices

Fall 2007					
Model	N	NPAR	NFI	CFI	RMSEA
One-Factor	247	21	0.833	0.855	0.148
Five-Factor Orthogonal	247	25	0.666	0.683	0.219
Five-Factor Oblique	247	25	0.890	0.912	0.119
Second-Order	247	26	0.890	0.912	0.119
Spring 2008					
Model	N	NPAR	NFI	CFI	RMSEA
One-Factor	277	21	0.871	0.891	0.131
Five-Factor Orthogonal	277	25	0.655	0.669	0.229
Five-Factor Oblique	277	25	0.915	0.935	0.105
Second-Order	277	26	0.910	0.929	0.109

Note. All chi-square values are significant at α = .001.

Table 41

Internship, Practica, and Clinical Course Form Pattern and Structure Coefficients

		Fall 2007		Spring 2008	
Factor	Question	Pattern Coefficient	Structure Coefficient	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1	0.513	0.802	0.357	0.829
	2	0.566	0.914	0.406	0.910
	3	0.497	0.884	0.415	0.916
	4	0.465	0.833	0.389	0.867
Enthusiasm and Intellectual Stimulation	5	0.440	0.914	0.366	0.865
	6	0.445	0.961	0.420	0.924
	7	0.478	0.921	0.452	0.933
	8	0.485	0.889	0.452	0.893
Rapport and Respect	9	0.480	0.912	0.437	0.909
	10	0.508	0.944	0.453	0.890
	11	0.450	0.867	0.392	0.787
	12	0.416	0.847	0.367	0.852
Feedback and Accessibility	13	0.431	0.914	0.368	0.897
	14	0.427	0.929	0.401	0.916
	15	0.473	0.903	0.429	0.932
	16	0.427	0.840	0.428	0.945
Student Perceptions of Learning	17	0.533	0.929	0.481	0.963
	18	0.491	0.878	0.454	0.914
	19	0.424	0.731	0.356	0.762
	20	0.418	0.766	0.382	0.809

Note. All coefficients are significant at the α = .05 level.

Table 42

Internship, Practica, and Clinical Course Form Pattern and Structure Coefficients for Perceived Instructional Effectiveness

	Fall 2007		Spring 2008	
First-Order Factor	Pattern Coefficient	Structure Coefficient	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1.278	0.947	1.414	0.959
Enthusiasm and Intellectual Stimulation	1.472	0.949	1.425	0.954
Rapport and Respect	1.365	0.969	1.390	0.977
Feedback and Accessibility	1.427	0.960	1.459	0.981
Student Perceptions of Learning	1.172	0.910	1.316	0.956

Note. All coefficients are significant at the α = .05 level.

The g-theory analysis shows that reliability is stable from the fall 2007 semester to the spring 2008 semester. There was a higher percentage of variation explained for teachers as opposed to courses for the spring 2008 semester but not for the fall 2007 semester (see Tables 43 and 44), indicating that making teacher based decisions is more appropriate than making course based decisions currently.

Table 43

Internship, Practica, and Clinical Course Form s:c:t x *i G-theory Analysis of*

	Fall 2007		Spring 2008	
Source of Variation	Estimate	Percent	Estimate	Percent
t	0.072	11.320%	0.086	18.145%
i	0.003	0.398%	0.002	0.438%
c:t	0.000	0.000%	0.000	0.000%
t *x* i	0.022	3.487%	0.005	0.955%
s:c:t	0.366	57.238%	0.275	57.833%
c *x* i:t	0.000	0.000%	0.002	0.518%
s:c:t *x* i, e	0.176	27.557%	0.105	22.111%

Table 44
Internship, Practica, and Clinical Course Form s:t:c x *i G-theory Analysis*

	Fall 2007		Spring 2008	
Source of Variation	Estimate	Percent	Estimate	Percent
c	0.078	11.981%	0.000	0.000%
i	0.002	0.369%	0.002	0.441%
t:c	0.000	0.000%	0.083	17.424%
c *x* i	0.007	1.072%	0.002	0.381%
s:t:c	0.353	54.317%	0.278	58.311%
t *x* i:c	0.008	1.245%	0.004	0.887%
s:t:c *x* i, e	0.202	31.016%	0.108	22.557%

The g coefficients for different scenarios based on the combination of number of courses taught and number of students in each class were also consistent from the fall 2007 semester to the spring 2008 semester (see Table 45). These g coefficients moderate at the worst but overall excellent for most scenarios.

Table 45
Internship, Practica, and Clinical Course Form Estimated s:c:t x *i G Coefficients for 20 Items*

	n_s									
	Fall 2007					Spring 2008				
n_c	5	10	15	20	25	5	10	15	20	25
1	0.488	0.652	0.735	0.785	0.818	0.605	0.753	0.819	0.857	0.882
2	0.652	0.785	0.842	0.873	0.894	0.753	0.858	0.900	0.922	0.936
3	0.735	0.842	0.885	0.908	0.922	0.820	0.900	0.930	0.946	0.956
4	0.785	0.873	0.908	0.926	0.937	0.858	0.922	0.946	0.958	0.966
5	0.818	0.894	0.922	0.937	0.946	0.883	0.936	0.956	0.966	0.972
6	0.842	0.908	0.932	0.945	0.952	0.900	0.946	0.963	0.971	0.976
7	0.860	0.918	0.939	0.950	0.957	0.913	0.953	0.967	0.975	0.979

An HLM model was estimated to further understand how the variance components of the g-theory analysis can be disentangled into support for construct validity or support against construct validity. The random base model (see Table 46) for *Perceived Instructional Effectiveness* yielded a reliability between sections of 0.010 and a reliability across teachers of 0.276, which are poor. These results were much lower than the g-theory results. The estimates of between-section and within-section variability of *Perceived Instructional Effectiveness* were 0.372 and 92.057 respectively, leading to an intraclass correlation (proportion of variance due to between-section differences) of 0.004 (see below).

Table 46

Internship, Practica, and Clinical Course Form Random Base Model: Random Effects

Random Effect	Variance	df	Chi-Square	P-Value
	Level-1 and Level-2			
Between Section	0.372	23	15.377	>.500
Within Section	92.057			
	Level-3			
Between Teachers	12.758	69	99.521	.010

$$\rho = \frac{0.37194}{0.37194 + 92.05714} = 0.004$$

This intraclass correlation indicates that 0.4% of the variance in ratings is between sections, without any predictor variables. The between-teacher variance was statistically significant, so higher-level models can explain more differences. Unfortunately, there is not enough data to estimate any higher-level models.

Laboratory Course Form

The Laboratory Course Form has excellent model fit (see Table 47). The NFI and CFI statistics meet their ideal cutoff values in the fall 2007 and spring 2008 semesters. This is strong evidence for construct validity.

Table 47
Laboratory Course Form Goodness-of-fit Indices

Model	N	NPAR	NFI	CFI	RMSEA
Spring 2007					
One-Factor	376	21	0.856	0.875	0.122
Five-Factor Orthogonal	376	25	0.671	0.685	0.193
Five-Factor Oblique	376	25	0.922	0.941	0.086
Second-Order	376	26	0.919	0.938	0.088
Fall 2007					
Model	N	NPAR	NFI	CFI	RMSEA
One-Factor	1002	21	0.880	0.888	0.109
Five-Factor Orthogonal	1002	25	0.677	0.683	0.184
Five-Factor Oblique	1002	25	0.946	0.955	0.072
Second-Order	1002	26	0.943	0.952	0.074
Spring 2008					
Model	N	NPAR	NFI	CFI	RMSEA
One-Factor	1086	21	0.886	0.893	0.105
Five-Factor Orthogonal	1086	25	0.681	0.687	0.180
Five-Factor Oblique	1086	25	0.955	0.963	0.064
Second-Order	1086	26	0.952	0.960	0.067

Note. All chi-square values are significant at α = .001.

Each question loaded positively and significantly at the alpha = .05 level on their proposed first-order factors for each semester (see Table 48). This offers further evidence for construct validity and suggests that this form does not need to be changed. Also, each first-order factor's loading on *Perceived Instructional Effectiveness* was also positive and significant, again showing evidence for construct validity (see Table 49).

Table 48

Laboratory Course Form Pattern and Structure Coefficients

Spring 2007

Factor	Question	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1	0.519	0.847
	2	0.471	0.802
	3	0.593	0.862
	4	0.578	0.893
Enthusiasm and Intellectual Stimulation	5	0.409	0.644
	6	0.591	0.904
	7	0.603	0.927
	8	0.508	0.792
Rapport and Respect	9	0.462	0.717
	10	0.568	0.675
	11	0.613	0.864
	12	0.664	0.819
Feedback and Accessibility	13	0.623	0.854
	14	0.646	0.922
	15	0.496	0.867
	16	0.620	0.889
Student Perceptions of Learning	17	0.615	0.914
	18	0.630	0.819
	19	0.605	0.864
	20	0.613	0.909

Fall 2007

Factor	Question	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1	0.480	0.841
	2	0.416	0.768
	3	0.548	0.850
	4	0.491	0.839
Enthusiasm and Intellectual Stimulation	5	0.465	0.709
	6	0.502	0.839
	7	0.543	0.891
	8	0.535	0.832
Rapport and Respect	9	0.416	0.718
	10	0.539	0.611
	11	0.562	0.847
	12	0.594	0.781
Feedback and Accessibility	13	0.416	0.789
	14	0.528	0.881
	15	0.420	0.822
	16	0.527	0.896
Student Perceptions of Learning	17	0.499	0.920
	18	0.527	0.858
	19	0.462	0.837
	20	0.504	0.928

Table 48 (Continued)

Spring 2008			
Factor	Question	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1	0.375	0.797
	2	0.395	0.834
	3	0.456	0.862
	4	0.423	0.830
Enthusiasm and Intellectual Stimulation	5	0.414	0.743
	6	0.388	0.773
	7	0.473	0.881
	8	0.479	0.827
Rapport and Respect	9	0.345	0.695
	10	0.456	0.671
	11	0.466	0.849
	12	0.469	0.805
Feedback and Accessibility	13	0.402	0.774
	14	0.467	0.885
	15	0.399	0.831
	16	0.470	0.876
Student Perceptions of Learning	17	0.418	0.915
	18	0.463	0.874
	19	0.381	0.828
	20	0.417	0.905

Note. All coefficients are significant at the α = .05 level.

Table 49

Laboratory Course Form Pattern and Structure Coefficients for Perceived Instructional Effectiveness

Spring 2007		
First-Order Factor	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1.503	0.924
Enthusiasm and Intellectual Stimulation	1.621	0.955
Rapport and Respect	1.135	0.914
Feedback and Accessibility	1.479	0.908
Student Perceptions of Learning	1.487	0.974
Fall 2007		
First-Order Factor	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1.280	0.903
Enthusiasm and Intellectual Stimulation	1.373	0.971
Rapport and Respect	1.041	0.935
Feedback and Accessibility	1.389	0.919
Student Perceptions of Learning	1.471	0.949
Spring 2008		
First-Order Factor	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1.149	0.887
Enthusiasm and Intellectual Stimulation	1.241	0.990
Rapport and Respect	1.019	0.932
Feedback and Accessibility	1.205	0.912
Student Perceptions of Learning	1.347	0.946

Note. All coefficients are significant at the $\alpha = .05$ level.

The g-theory analysis shows a lack of reliability in the spring 2007 semester. There was initially a higher reliability across teachers as opposed to across courses, indicating that students used their perception of the teacher more than the course in their responses (see Tables 50 and 51).

Table 50

Laboratory Course Form s:c:t x *i G-theory Analysis*

	Spring 2007		Fall 2007		Spring 2008	
Source of Variation	Estimate	Percent	Estimate	Percent	Estimate	Percent
t	0.099	8.366%	0.047	6.326%	0.029	6.511%
i	0.030	2.541%	0.012	1.566%	0.005	1.028%
c:t	0.000	0.000%	0.042	5.630%	0.024	5.327%
t *x* i	0.026	2.225%	0.019	2.524%	0.012	2.621%
s:c:t	0.626	52.819%	0.337	45.530%	0.214	47.692%
c *x* i:t	0.024	2.038%	0.007	0.989%	0.006	1.253%
s:c:t *x* i, e	0.379	32.011%	0.277	37.435%	0.159	35.568%

Table 51

Laboratory Course Form s:t:c x *i G-theory Analysis*

	Spring 2007		Fall 2007		Spring 2008	
Source of Variation	Estimate	Percent	Estimate	Percent	Estimate	Percent
c	0.056	4.758%	0.000	0.000%	0.010	2.250%
i	0.028	2.378%	0.011	1.527%	0.005	1.029%
t:c	0.062	5.217%	0.072	9.634%	0.033	7.406%
c *x* i	0.024	2.024%	0.012	1.567%	0.005	1.225%
s:t:c	0.603	51.081%	0.363	48.324%	0.223	49.866%
t *x* i:c	0.021	1.755%	0.011	1.443%	0.011	2.397%
s:t:c *x* i, e	0.387	32.786%	0.282	37.505%	0.161	35.826%

Table 52

Laboratory Course Form Estimated s:c:t x *i G Coefficients for 20 Items*

	Spring 2007				
	n_s				
n_c	5	10	15	20	25
1	0.430	0.597	0.685	0.740	0.778
2	0.599	0.744	0.809	0.846	0.870
3	0.689	0.810	0.861	0.888	0.906
4	0.745	0.848	0.889	0.911	0.925
5	0.784	0.873	0.907	0.925	0.937
6	0.812	0.890	0.919	0.935	0.945
7	0.833	0.903	0.929	0.942	0.950
	Fall 2007				
	n_s				
n_c	5	10	15	20	25
1	0.293	0.375	0.414	0.436	0.451
2	0.451	0.542	0.582	0.604	0.618
3	0.550	0.637	0.673	0.692	0.705
4	0.618	0.698	0.730	0.747	0.758
5	0.667	0.741	0.770	0.785	0.794
6	0.705	0.773	0.798	0.812	0.820
7	0.734	0.797	0.820	0.832	0.840
	Spring 2008				
	n_s				
n_c	5	10	15	20	25
1	0.297	0.383	0.425	0.449	0.465
2	0.456	0.551	0.593	0.616	0.631
3	0.555	0.645	0.683	0.703	0.716
4	0.622	0.706	0.739	0.756	0.767
5	0.671	0.748	0.777	0.793	0.802
6	0.709	0.778	0.805	0.819	0.827
7	0.738	0.802	0.826	0.838	0.846

An HLM model was estimated to further understand how the variance components of the g-theory analysis can be disentangled into support for construct validity or support against construct validity. The random base model (see Table 53) for *Perceived Instructional Effectiveness* yielded a reliability between sections of 0.355 and a reliability across teachers of

0.568. These reliabilities were moderate and consistent with the g-theory results. The estimates of between-section and within-section variability of *Perceived Instructional Effectiveness* were 0.372 and 92.057 respectively, leading to an intraclass correlation (proportion of variance due to between-section differences) of 0.084 (see below).

Table 53

Laboratory Course Form Random Base Model: Random Effects

Random Effect	Variance	df	Chi-Square	P-Value
	Level-1 and Level-2			
Between Section	7.641	87	14.613	<.001
Within Section	83.655			
	Level-3			
Between Teachers	14.399	45	133.459	<.001

$$\rho = \frac{7.64060}{7.64060 + 83.65456} = 0.084$$

This intraclass correlation indicates that 8.4% of the variance in ratings is between sections, without any predictor variables. The between-teacher variance was statistically significant, meaning that there are still differences that can be explained by higher-level models. Unfortunately, there was not enough data to estimate such models.

Studio-Performance Course Form

The Studio-Performance Course Form has good model fit (see Table 54). Oddly, the number of observations increased from the fall 2007 semester to the spring 2008 semester but the model fit decreased. This indicates that despite having moderate evidence for construct validity, there might be problems with this course form.

Table 54
Studio-Performance Course Form Goodness-of-fit Indices

Model	N	NPAR	NFI	CFI	RMSEA
Fall 2007					
One-Factor	347	21	0.822	0.838	0.151
Five-Factor Orthogonal	347	25	0.669	0.682	0.212
Five-Factor Oblique	347	25	0.902	0.919	0.110
Second-Order	347	26	0.901	0.918	0.111
Spring 2008					
Model	N	NPAR	NFI	CFI	RMSEA
One-Factor	514	21	0.726	0.735	0.198
Five-Factor Orthogonal	514	25	0.627	0.635	0.233
Five-Factor Oblique	514	25	0.808	0.817	0.170
Second-Order	514	26	0.805	0.814	0.171

Note. All chi-square values are significant at α = .001.

Each question loaded positively and significantly on its expected first-order factor (see Table 55). Upon comparing the structure coefficients within each factor, it appears that questions 1, 3, 9, 10, 13, and 14 might be good places for improvement. The first-order factors also loaded positively and significantly on *Perceived Instructional Effectiveness* (see Table 56), providing some further support for construct validity.

Table 55

Studio-Performance Course Form Pattern and Structure Coefficients

		Fall 2007		Spring 2008	
Factor	Question	Pattern Coefficient	Structure Coefficient	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1	0.401	0.795	0.196	0.747
	2	0.576	0.934	0.295	0.874
	3	0.590	0.908	0.290	0.788
	4	0.560	0.893	0.324	0.903
Enthusiasm and Intellectual Stimulation	5	0.424	0.847	0.348	0.904
	6	0.508	0.887	0.404	0.888
	7	0.533	0.934	0.394	0.884
	8	0.515	0.907	0.362	0.922
Rapport and Respect	9	0.678	0.948	0.312	0.760
	10	0.548	0.795	0.268	0.664
	11	0.621	0.768	0.488	0.900
	12	0.552	0.737	0.455	0.910
Feedback and Accessibility	13	0.547	0.781	0.420	0.829
	14	0.560	0.924	0.332	0.829
	15	0.559	0.870	0.426	0.874
	16	0.606	0.849	0.419	0.942
Student Perceptions of Learning	17	0.492	0.895	0.378	0.910
	18	0.484	0.928	0.349	0.888
	19	0.518	0.883	0.351	0.924
	20	0.454	0.913	0.378	0.929

Note. All coefficients are significant at the α = .05 level.

Table 56

Studio-Performance Course Form Pattern and Structure Coefficients for Perceived Instructional Effectiveness

	Fall 2007		Spring 2008	
First-Order Factor	Pattern Coefficient	Structure Coefficient	Pattern Coefficient	Structure Coefficient
Organization and Clarity	1.533	0.963	1.193	0.891
Enthusiasm and Intellectual Stimulation	1.365	0.926	1.298	0.943
Rapport and Respect	1.313	0.950	1.142	0.841
Feedback and Accessibility	1.384	0.972	1.180	0.928
Student Perceptions of Learning	1.289	0.890	1.309	0.943

Note. All coefficients are significant at the α = .05 level.

The g-theory analysis gives evidence of reliability in that the Studio-Performance Course Form is more reliable for making decisions about teachers than for about courses (see Tables 57 and 58). Also, the g coefficients for most combinations of number of courses and students show good reliability (see Table 59).

Table 57

Studio-Performance Course Form s:c:t x *i G-theory Analysis*

	Fall 2007		Spring 2008	
Source of Variation	Estimate	Percent	Estimate	Percent
t	0.195	20.910%	0.009	2.842%
i	0.009	0.973%	0.003	0.796%
c:t	0.000	0.000%	0.000	0.000%
t *x* i	0.040	4.247%	0.008	2.531%
s:c:t	0.431	46.354%	0.218	69.122%
c *x* i:t	0.016	1.706%	0.007	2.200%
s:c:t *x* i, e	0.240	25.809%	0.071	22.509%

Table 58

Studio-Performance Course Form s:t:c x *i G-theory Analysis*

	Fall 2007		Spring 2008	
Source of Variation	Estimate	Percent	Estimate	Percent
c	0.063	7.253%	0.000	0.000%
i	0.010	1.165%	0.003	0.832%
t:c	0.087	10.013%	0.007	2.340%
c *x* i	0.003	0.365%	0.001	0.302%
s:t:c	0.417	47.798%	0.220	69.568%
t *x* i:c	0.050	5.757%	0.011	3.518%
s:t:c *x* i, e	0.241	27.649%	0.074	23.441%

Table 59

Studio-Performance Course Form Estimated s:c:t x *i G Coefficients for 20 Items*

	n_s									
	Fall 2007					Spring 2008				
n_c	5	10	15	20	25	5	10	15	20	25
1	0.680	0.805	0.858	0.886	0.905	0.166	0.281	0.366	0.431	0.483
2	0.806	0.888	0.919	0.935	0.945	0.283	0.435	0.530	0.595	0.642
3	0.860	0.920	0.941	0.953	0.960	0.370	0.532	0.622	0.681	0.721
4	0.889	0.936	0.953	0.962	0.967	0.437	0.598	0.682	0.734	0.768
5	0.907	0.946	0.960	0.967	0.972	0.490	0.647	0.724	0.770	0.800
6	0.920	0.953	0.965	0.971	0.975	0.533	0.684	0.754	0.796	0.823
7	0.930	0.959	0.969	0.974	0.977	0.569	0.713	0.778	0.815	0.839

An HLM model was estimated to further understand how the variance components of the g-theory analysis can be disentangled into support for construct validity or support against construct validity. The random base model (see Table 60) for *Perceived Instructional Effectiveness* yielded a reliability between sections of 0.303 and a reliability across teachers of 0.685. These reliabilities were moderate and consistent with the g-theory results. The estimates of between-section and within-section variability of *Perceived Instructional Effectiveness* were 14.652 and 84.331 respectively, leading to an intraclass correlation (proportion of variance due to between-section differences) of 0.148 (see below).

Table 60

Studio-Performance Course Form Random Base Model: Random Effects

Random Effect	Variance	df	Chi-Square	P-Value
	Level-1 and Level-2			
Between Section	14.652	61	31.330	.008
Within Section	84.331			
	Level-3			
Between Teachers	54.295	38	8.197	<.001

$$\rho = \frac{14.65228}{14.65228 + 84.33085} = 0.148$$

This intraclass correlation indicates that 14.8% of the variance in ratings is between sections, without any predictor variables. The between-teacher variance was statistically significant, so higher-level models can explain more differences. Unfortunately, there are not enough data to estimate any higher-level models.

DISCUSSION

The reliability and validity varied among course forms. Overall, the second-order model was equivalent in fit to the five-factor oblique model, so all analyses were conducted on the second-order model. This model fit was better than the other models for each course form. This model had the benefit of having a higher-order factor that summarized each course form as a whole which was labeled *Perceived Instructional Effectiveness*. Overall, the structure coefficients for each semester and each course form were highly homogeneous. Each item's structure coefficient was high and positive, implying that each item measured some positive component of its respected factor and *Perceived Instructional Effectiveness*.

Beyond using confirmatory factor analysis (CFA) to investigate construct validity through analyzing model fit, CFA was also used to give recommendations to which questions were troublesome and could be changed to improve construct validity. Overall, the Laboratory Course Form, Online Course Form, and Standard Lecture Course Form showed excellent model fit. The Activities Course Form, Internship, Practica, and Clinical Course Form, Seminar Course Form, and Studio-Performance Course Form all showed good model fit. The Independent Research Course Form showed very poor model fit. Also, the *Enthusiasm and Intellectual Stimulation* factor played a large role in both the Activities Course Form and the Independent Research Course Form. This suggests that enthusiastic teachers with interesting activities and who are interested in their students' projects are better teachers for activities courses and independent research courses. Questions 2, 7, 8, 12, and 13 on the Activities Course Form are troublesome because they do not explain much of their proposed factor. Questions 1, 2, 5, 8, 9, 10, and 13 are troublesome on the Independent Research Course Form. Changing questions 2 and 7 might improve validity on the Seminar Course Form. Finally, questions 1, 3, 9, 10, 13, and 14 are problematic on the Studio-Performance Course Form and could use improvement.

The amount of variance from the generalizability theory (g-theory) analyses for the Laboratory Course Form, Online Course Form, Seminar Course Form, Standard Lecture Course Form, and Studio-Performance Course Form were higher for teachers than for courses, giving evidence of reliability. This implies that making decisions about teachers is more appropriate than making decisions about courses for those course forms. The Independent Research Course Form was the opposite in that the amount of variance was higher for courses than for teachers. The variability was virtually nothing for teachers on this form. It is possible that this is because of the higher amount of independent learning in these courses, or this could be a general reliability issue. This was also the case with the Activities Course Form, which yielded a higher amount of variance for courses than for teachers in the spring 2008 semester. The Internship, Practica, and Clinical Course Form also yielded a higher amount of variance for courses than for teachers, but this changed in the spring 2008 semester showing an increase in reliability. Overall, making decisions on the Activities Course Form and Independent Research Course Form should be done with caution. The generalizability coefficients (g coefficients) for different scenarios of number of courses taught and number of students in each course taught showed moderate to excellent reliability on all course forms except the Independent Research, and Laboratory course forms. These good g coefficients are consistent with the literature (Gillmore, Kane, & Naccarato, 1978; Smith, 1979). The reliability of those forms needs to be addressed and improved.

It is difficult to determine how much of the variance in each facet is error due to other variables that could be biases or evidence of validity. Therefore, hierarchical linear modeling (HLM) was used to determine what percent of *Perceived Instructional Effectiveness* was due to biases and what variables were creating this bias. Random base models, which are one-way random effects models that are unconditional in that they did not specify any predictor variables, show that potential biases explain only a small amount of ratings. The proportion of variance in ratings accounted for between section in these random base models ranged from 0.003% to 14.8%. The Online and Seminar Course Forms were able to be estimated beyond the random base model to a level-1 model. The amount of between-section variance that can be explained by student measures was estimated to be 9.6% on the Online Course Form and 7.5% on the Seminar Course Form. The amount of between-teacher variance that can be explained by student measures was estimated to be 9.7% on the Online Course Form.

The Standard Lecture Course Form was the only course form to have a level-3 model. Table 61 outlines the percentage of variance explained between sections and between teachers by student characteristics, course characteristics, and teacher characteristics. Since many hypotheses failed to show support for biases, it is certain that biases do not account for the entire 28.8% of ratings that are explained by student, course, and teacher measures.

Table 61

Percentage of Between-Section and Between-Teacher Variance Explained by Student, Course, and Teacher Characteristics

Measure	% of Between-Section Variance	% of Between-Teacher Variance
Student Measures	0.2	21.0
Course Measures	2.8	6.8
Teacher Measures	N/A	1.0
Total	3.0	28.8

Because the Online Course Form and Seminar Course Form had enough data to estimate a level-1 model, some potential biases were investigated. Variables that showed up as biases on the Online Course Form were the reason for taking the course, interest level, and attempted hours. It is possible that students who have a higher interest in the course also have higher expectations, and that higher level of standard they expect from the teacher ultimately leads to lower ratings. Higher interest level also negatively impacted ratings on the Seminar Course Form. The number of hours spent outside of class seemed to negatively influence students' perceptions of how much they learned in online courses and overall course workload seemed to do the same for seminar courses.

Because the Standard Lecture Course Form had enough data to estimate a level-3 model, all potential biases were investigated. Because of this, this was the only form that allowed the investigation of the expected grade hypotheses. For students' expected grades, the "grading-leniency hypothesis" and "course specific motivation hypothesis" were both accepted. This might signify that students at Western Carolina University (WCU) prefer surface approach learning, or learning that is well-organized and allows for preparing for tests through passive listening. These students may be more concerned with having a good time and having easier teachers throughout their college career. It is likely that grade inflation is giving support for the "grading-leniency hypothesis," since easy grading seems to be positively biasing ratings. Also, students' motivation appears to change from course to course and may be attributed to instruction. Teachers can get better ratings in their courses by improving their students' motivation for those courses. Overall, it appears that the current study does not support the large body of literature suggesting that the "validity hypothesis" and "students' characteristics hypothesis" have the largest body of evidence. Expected grades appear to bias ratings at WCU.

Fortunately, as can be seen in Table 61, all potential biases do not explain much of the ratings. Also of interest was the finding that male students tend to give lower ratings and female teachers tend to receive higher ratings. Overall course workload tended to influence students' rapport and respect for their teachers negatively as well as their perception of how accessible the teacher is for feedback and the quality of the feedback. Workload outside of the class in excess of 11 hours also negatively influenced ratings for rapport and respect and students' perceptions of their learning, implying that this workload is not fair and appropriate. Students' perceptions of their learning in the course were negatively associated with the amount of writing in the course.

The biggest limitation is the impossibility to determine the reasons for reliability and validity increases and decreases. Between the fall 2007 and spring 2008 semesters, more policies and procedures were implemented and the number of response options was reduced from five to four. This study could not control for this large number of changes. Also, the low response rate was also a limitation. It is likely that increasing the response rate in the future will yield better reliability and validity. It should also be noted that some portions of the analysis for the Activities Course Form, Independent Research Course Form, Internship, Practica, and Clinical Course Form, Laboratory Course Form, and Studio-Performance Course Form are not particularly useful due to the low number of responses. Improving the response rate and changing some of the questions and factors on the forms mentioned above can improve reliability and validity in the future and increase the number of responses, thereby yielding a better analysis. Another limitation was the inclusion of the 11 Liberal Studies questions. These questions were confusing to students and may have lowered the reliability and validity of each course form. Ensuring that questions such as those are not included on the Student Assessment of Instruction (SAI) would improve reliability and validity.

It is clear from these limitations that student ratings are not truly objective, but also that teachers can be one source of bias for their own ratings. It is also clear that there are some cases in which students are essentially consumers who want to be cheated. It would be beneficial to conduct future research to improve student ratings and use them to improve the quality of education. Future research should attempt to determine whether or not low response rates lead to non-response error. Also, this study could not determine the impact of the transmission/teacher focused approach and the conceptual change/student focused approach to teaching on students’ evaluations. Future research also needs to investigate whether or not summer courses are as valid as regular semester courses, whether or not low enrolled courses are as valid as non-low enrolled courses, whether or not non-full semester courses are as valid as full semester courses, and whether or not graduate students give more valid evaluations than undergraduate students. Finally, better analysis of expected grades, such as a path analysis, might be better to test each expected grade hypothesis. Addressing grade inflation and conducting research across multiple institutions to account for institutional type would be very beneficial for improving students’ assessments of instruction.

Overall, the Independent Research Course Form had no evidence of reliability and validity, whereas the Online Course Form had some evidence of validity but little evidence of reliability. All other course forms had moderate to excellent reliability and validity. If we extrapolate the results from the Standard Lecture Course Form and apply them to the other course forms, which could not be estimated at level-3, we find that at the very most 28.8% of ratings can be accounted for by potential biases. Despite this, model fit for the Standard Lecture Course Form was excellent implying that these biases have a small effect on reducing construct validity. This in conjunction with the good reliability and validity results suggest that the SAI can be used appropriately for both formative and summative purposes. It appears that since the Online, Seminar, and Standard Lecture course forms had good validity, creating separate evaluation instruments for each instructional method is beneficial.

REFERENCES

Abrami, P. C., & d'Apollonia, S. (1991). Multidimensional student's evaluations of teaching effectiveness-generalizability of "N = 1" research: Comment on Marsh. *Journal of Educational Psychology, 83* (3), 411-415.

Aleamoni, L. M. (1981). Student ratings of instruction. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 110-145). Beverly Hills, CA: Sage Publications.

Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley, L., et al. (2004). Student evaluation of college teaching. *College Teaching, 52* (4), 134-141.

Archer, T. (2003). Web-based surveys. *Journal of Extension* [On-line], *41* (4). Available at: http://www.joe.org/joe/2003august/tt6.shtml.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin , 88* (3), 588-606.

Carini, R. M., Hayek, J. C., Kuh, G. D., Kennedy, J. M., & Ouimet, J. A. (2003). College student responses to web and paper surveys: Does mode matter? *Research in Higher Education, 44* (1), 1-19.

Cohen, P. A., & McKeachie, W. J. (1980). The role of colleagues in the evaluation of college teaching. *Improving College and University Teaching, 28* (4), 147-154.

Conrad, C. F., & Pratt, A. M. (1985). Designing for quality. *Journal of Higher Education, 56* (6), 601-622.

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64* (3), 391-418.

Cui, W. (2003). *Reducing error in mail surveys. ERIC Digest.* Retrieved February 20, 2008, from ERIC database.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52* (11), 1198-1208.

Ellis, L., Burke, D. M., Lomire, P., & McCormack, D. R. (2003). Student grades and average ratings of instructional quality: The need for adjustment. *The Journal of Educational Research, 97* (1), 35-4.

Ethington, C. A. (1997). A HLM approach to studying college effects. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 12). New York, NY: Agathon Press.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education, 9*, 199-242.

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* (6), 583-645.

Franklin, J., & Theall, M. (1992, April 20-24). *Disciplinary differences: Instructional goals and activities, measures of student performance, and student ratings of instruction*. Paper presented at the Annual Conference of the American Educaitonal Research Association, San Francisco, California.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement , 15* (1), 1-13.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52* (11), 1209-1217.

Groves, R. M., Couper, M. P., Presser, S., Singer, E., Tourangeau, R., Acosta, G. P., et al. (2006). Experiments in producing nonresponse bias. *Public Opinion Quarterly, 70* (5), 720-736.

Harrison, P. D., Douglas, D. K., & Burdsal, C. A. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education, 45* (3), 311-323.

Helgeson, J. G., Voss, K. E., & Terpening, W. D. (2002). Determinants of mail-survey response: Survey design factors and respondent factors. *Psychology & Marketing, 19* (3), 303-328.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6* (1), 1-55.

Kraiger, K., & Teachout, M. S. (1990). G-theory as a construct-related evidence of the validity of job performance ratings. *Human Performance, 3* (1), 19-35.

Kyriakides, L., Demetriou, D., & Charalambous, C. (2006). Generating criteria for evaluating teachers through teacher effectiveness research. *Educational Research, 48* (1), 1-2.

MacCallum, R. C. (1995). Model specification: Procedures, strategies, and related issues. In R. H. Hoyle (Ed.), *Structural equation modeling* (pp. 16-36). Thousand Oaks, CA: Sage Publications, Inc.

Macdonald, J. B. (1980a). Evaluation of teaching: Purpose, context, and problems. In W. R. Duckett, W. J. Gephart, R. B. Ingle, & M. R. Carroll (Eds.), *Planning for the evaluation of teaching* (pp. 13-26). Bloomington, IN: Phi Delta Kappa.

Macdonald, J. B. (1980b). Planning for the evaluation of teaching. In W. R. Duckett, W. J. Gephart, R. B. Ingle, & M. R. Carroll (Eds.), *Planning for the evaluation of teaching* (pp. 2-12). Bloomington, IN: Phi Delta Kappa.

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

Marsh, H. W. (1991). A multidimensional perspective on students' evaluations of teaching effectiveness: Reply to Abrami and d'Apollonia. *Journal of Educational Psychology, 83* (3), 416-421.

Marsh, H. W. (1992, April 20-24). *A longitudinal perspective of students' evaluations of university teaching: Ratings of the same teachers over a 13-year period.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California.

Marsh, H. W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology, 99* (4), 775-79.

Marsh, H. W., Hau, K., Chung, C., & Siu, T. L. (1997). Students' evaluations of university teaching: Chinese version of the Students' Evaluations of Educational Quality Instrument. *Journal of Educational Psychology, 89* (3), 568-572.

Marsh, H., & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. *American Educational Research Journal , 21* (2), 341-366.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52* (11), 1187-1197.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52* (11), 1218-1225.

Miller, A. D., & Murdock, T. B. (2007). Modeling latent true scores to determine the utility of aggregate student perceptions as classroom indicators in HLM: The case of classroom goal structures. *Contemporary Educational Psychology, 32* (1), 83-104.

Patrick, W. J. (2001). Estimating first-year student attrition rates: An application of multilevel modeling using categorical variables. *Research in Higher Education, 42* (2), 151-17.

Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education, 44* (4), 409-432.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53*, 901-912.

Schrodt, P., Turman, P. D., & Soliz, J. (2006). Perceived understanding as a mediator of perceived teacher confirmation and students' ratings of instruction. *Communication Education, 55* (4), 370-388.

Scriven, M. (1980). A different approach to teacher evaluation. In W. R. Duckett, W. J. Gephart, R. B. Ingle, & M. R. Carroll (Eds.), *Planning for the evaluation of teaching* (pp. 60-72). Bloomington, IN: Phi Delta Kappa.

Seldin, P. (1993). The use and abuse of student ratings of professors. *The Chronicle of Higher Education, 39* (46), 4.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology, 34* (2), 133-166.

Smith, P. L. (1979). The generalizability of student ratings of courses: Asking the right questions. *Journal of Educational Measurement*, *16* (2), 77-87.

Spencer, K. J., & Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. *Assessment & Evaluation in Higher Education, 27* (5), 397-409.

Stronge, J. H. (1997). Improving schools through teacher evaluation. In J. H. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practices* (pp. 1-23). Thousand Oaks, CA: Corwin Press, Inc.

Stronge, J. H., & Ostrander, L. P. (1997). Client surveys in teacher evaluation. In J. H. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practices* (pp. 129-161). Thousand Oaks, CA: Corwin Press, Inc.

Suen, H. K., & Lie, P. (2007). Classical versus G-theory of measurement. Educational Measurement, *4*, 3-2.

Theall, M., & Franklin, J. (2000). Creating responsive student ratings systems to improve evaluation practice. *New Directions for Teaching and Learning* , *83*, 95-107.

Theall, M., & Franklin, J. (2001). Using technology to facilitate evaluation. *New Directions for Teaching and Learning* , *88,* 41-5.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association.

Thorpe, S. W. (2002, June 2-5). *Online student evaluation of instruction: An investigation of non-response bias.* Paper presented at the Annual Forum for the Association for Institutional Research, Toronto, Ontario, Canada.

Travers, R. M. (1981). Criteria of good teaching. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 14-22). Beverly Hills, CA: Sage Publications.

Weiss, D. J. (1971). Further considerations in applications of factor analysis. *Journal of Counseling Psychology, 18* (1), 85-92.

Weiss, D. J., & Davison, M. L. (1981). Test theory and methods. *Annual Review of Psychology, 32*, 629-658.

Wheeler, P. H., & Scriven, M. (1997). Building the foundation: Teacher roles and responsibilities. In J. H. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practices* (pp. 27-58). Thousand Oaks, CA: Corwin Press, Inc.

Willcoxson, L. (1998). The impact of academics' learning and teaching preferences on their teaching practices: A pilot study. *Studies in Higher Education, 23* (1), 59-7.

Wilson, R. C., Dienst, E. R., & Watson, N. L. (1973). Characteristics of effective college teachers as perceived by their colleagues. *Journal of Educational Measurement, 10* (1), 31-37.

Wotruba, T. R., & Wright, P. L. (1975). How to develop a teacher-rating instrument. *Journal of Higher Education, 46* (6), 653-663.

Zhang, L. (2001). Approaches and thinking styles in teaching. *Journal of Psychology, 135* (5), 547-562.

Appendices

Appendix A

SAI Course Forms

Student Assessment of Instruction: Activities Course Form

Organization and Clarity

1. My teacher is well prepared for class meetings.
2. My teacher explains the course activities clearly.
3. My teacher provides good demonstration of the course activities.
4. My teacher provides sufficient time to practice the activity skills.

Enthusiasm and Intellectual Stimulation

5. My teacher is enthusiastic about teaching this course.
6. My teacher makes the course activities interesting.
7. My teacher motivates me to participate in these activities outside of class.
8. My teacher motivates student involvement in course activities.

Rapport and Respect

9. My teacher develops a close rapport with the class.
10. My teacher is regularly available for consultation.
11. My teacher deals with students as individuals.
12. My teacher is impartial in dealing with students.

Feedback and Accessibility

13. My teacher assigns grades fairly.
14. My teacher gives helpful feedback.
15. Assessment methods accurately measure what the teacher expects of me.
16. My teacher spends sufficient time to develop my activity skills.

Student Perceptions of Learning

17. My teacher provides practice that helps me learn the course activities.
18. My teacher encourages me to value the course activities.
19. My teacher promotes my grasp of important principles of the course activities.
20. My teacher helps me improve my activity skills.

Student Assessment of Instruction: Independent Research Form

Organization and Clarity

1. My research advisor helps me know what is expected of me in my project.
2. My research advisor gives me appropriate help with difficult aspects of my project.
3. My research advisor helps me organize my project.
4. My research advisor helps me keep my project on schedule.

Enthusiasm and Intellectual Stimulation

5. My research advisor is enthusiastic about my project.
6. My research advisor motivates me to complete my project.
7. My research advisor stimulates my thinking.
8. This experience makes me want to do independent research in the future.

Rapport and Respect

9. It is easy to talk with my research advisor about my project.
10. My research advisor and I have a good working relationship.
11. My research advisor respects my questions about the subject matter.
12. My research advisor respects my views of the project.

Feedback and Accessibility

13. My research advisor is readily available for consultation.
14. I have sufficient meetings with my research advisor on my project.
15. My research advisor lets me work independently in appropriate ways.
16. My research advisor provides me with helpful feedback on my project.

Student Perceptions of Learning

17. My research advisor has advanced my knowledge in the area of my project.
18. My research advisor helps me work more independently.
19. My research advisor stimulates my curiosity about my project.
20. My research advisor encourages me to value new viewpoints related to my project.

Student Assessment of Instruction: Internship, Practica and Clinical Course Form

Organization and Clarity

1. My teacher makes the requirements for this course clear.
2. My teacher coordinates interactions with work-site staff to my benefit.
3. Observation and supervision of my work are effective.
4. My teacher answers questions appropriately.

Enthusiasm and Intellectual Stimulation

5. My teacher is enthusiastic about supervising this course.
6. My teacher promotes my engagement in this course.
7. My teacher stimulates my thinking through this experience.
8. My teacher is fully engaged in supervising my work.

Rapport and Respect

9. My teacher actively helps me with course-related problems.
10. My teacher meets my needs for consultation.
11. My teacher respects opinions different from his or her own.
12. My teacher conveys appreciation to work-site staff.

Feedback and Accessibility

13. My teacher evaluates my performance fairly.
14. My teacher collects sufficient evidence for valid grading.
15. My teacher offers specific advice to promote improvement.
16. My teacher integrates constructive feedback from work-site staff.

Student Perceptions of Learning

17. My teacher enhances my ability to solve actual problems in my discipline.
18. My teacher enables me to connect theory with practice.
19. I am learning a lot from this course.
20. The course setting is conducive to my learning.

Student Assessment of Instruction: Laboratory Course Form

Organization and Clarity

1. My lab teacher is well prepared.
2. I know what is expected of me in this course.
3. My lab teacher explains the lab procedures clearly.
4. My lab teacher promotes good use of laboratory time.

Enthusiasm and Intellectual Stimulation

5. The lab assignments are interesting.
6. My lab teacher is enthusiastic about teaching this class.
7. My lab teacher motivates me to do well in the laboratory.
8. My lab teacher reinforces what I have learned in the lecture.

Rapport and Respect

9. My lab teacher insists that we all follow safety procedures.
10. My lab teacher is impartial in dealing with students.
11. My lab teacher respects student questions about the subject matter.
12. My lab teacher is regularly available for consultation.

Feedback and Accessibility

13. My lab teacher evaluates my work promptly.
14. My lab teacher provides helpful feedback on my progress.
15. Evaluations in this laboratory course are fair.
16. My lab teacher offers specific advice to promote improvements.

Student Perceptions of Learning

17. My lab teacher advances my knowledge in this lab section.
18. My lab teacher makes me more curious about the subject matter.
19. My lab teacher encourages me to work better with others in this course.
20. My teacher helps me learn important techniques in this course.

Student Assessment of Instruction: Online Course Form

Organization and clarity

1. My teacher provides clear guidelines for the work required in this course.
2. My teacher spaces assignments so they are due at reasonable intervals.
3. My teacher arranges assignments so they build on previous learning.
4. My teacher is flexible when there are disruptions in online access.

Enthusiasm and intellectual stimulation

5. My teacher stimulates my thinking.
6. My teacher helps me push my learning to new levels.
7. My teacher encourages open discussions.
8. My teacher helps keep me engaged in this course.

Rapport and respect

9. My teacher fosters mutual respect among students.
10. My teacher provides a safe environment for communication.
11. I am learning to value new viewpoints in this course.
12. My teacher fosters collaboration effectively.

Feedback and accessibility

13. My teacher gives feedback promptly enough to benefit me.
14. My teacher is clear about when she or he is accessible online.
15. Grades are assigned fairly.
16. Grading methods accurately measure what I am learning in this course.

Student perceptions of learning

17. My teacher promotes my understanding of important conceptual themes.
18. My teacher encourages students to learn from each other.
19. My teacher provides varied learning opportunities.
20. My teacher enhances my ability to communicate effectively about course subjects.

Student Assessment of Instruction: Seminar Course Form

Organization and Clarity

1. My teacher is well prepared for class meetings.
2. I know what is expected of me in this course.
3. My teacher poses questions that stimulate discussion.
4. The discussion sessions are well organized.

Enthusiasm and Intellectual Stimulation

5. My teacher makes me feel engaged in this class.
6. Discussions in this class are stimulating.
7. The teacher is enthusiastic about teaching this course.
8. My teacher motivates me to do well in this course.

Rapport and Respect

9. My teacher has a close rapport with the class.
10. My teacher is impartial in dealing with students.
11. My teacher respects student questions about the subject matter.
12. My teacher respects opinions different from his or her own.

Feedback and Accessibility

13. My teacher is readily available for consultation.
14. Evaluations in this course are fair.
15. Feedback from the teacher clearly indicates my standing in this course.
16. My teacher offers specific advice to promote improvements.

Student Perceptions of Learning

17. My teacher advances my knowledge of course content.
18. My teacher helps me to learn to work better with other students.
19. My teacher enhances my capacity to communicate effectively about the course content.
20. My teacher encourages me to value new viewpoints related to course content.

Student Assessment of Instruction: Standard Course Form

Organization and Clarity

1. My teacher is well prepared for class meetings.
2. My teacher explains the subject matter clearly.
3. My teacher clearly communicates course goals and objectives.
4. My teacher answers questions appropriately.

Enthusiasm and Intellectual Stimulation

5. My teacher is enthusiastic about teaching this course.
6. My teacher presents the subject in an interesting manner.
7. My teacher stimulates my thinking.
8. My teacher motivates me to do my best work.

Rapport and Respect

9. My teacher helps students sufficiently with course-related issues.
10. My teacher is regularly available for consultation.
11. My teacher is impartial in dealing with students.
12. My teacher respects opinions different from his or her own.

Feedback and Accessibility

13. Assessment methods accurately assess what I have learned in this course.
14. Grades are assigned fairly.
15. The basis for assigning grades is clearly explained.
16. The teacher provides feedback on my progress in the course on a regular basis.

Student Perceptions of Learning

17. My teacher advances my knowledge of course content.
18. My teacher promotes my understanding of important conceptual themes.
19. My teacher enhances my capacity to communicate effectively about the course subject matter.
20. My teacher encourages me to value new viewpoints related to the course.

Student Assessment of Instruction: Studio-Performance Course Form

Organization and Clarity

1. My teacher is well prepared for class meetings.
2. My teacher explains the subject clearly.
3. My teacher answers questions carefully and precisely.
4. My teacher gives clear assignments.

Enthusiasm and Intellectual Stimulation

5. My teacher is enthusiastic about teaching this course.
6. My teacher stimulates my creative expression.
7. My teacher motivates me to do my best work.
8. My teacher motivates student involvement.

Rapport and Respect

9. My teacher provides students sufficient help with course-related issues.
10. My teacher is regularly available for consultation.
11. My teacher is fair and impartial in dealing with students.
12. My teacher accepts opinions different from his or her own.

Feedback and Accessibility

13. My teacher provides sufficient individual instruction to me.
14. Assessment methods accurately measure what the teacher expects of me.
15. The basis for assessing my performance is clearly explained.
16. My teacher provides feedback promptly enough to benefit me.

Student Perceptions of Learning

17. I have gained a good grasp of concepts and techniques in this course.
18. I have enhanced my creative ability in this course.
19. I have learned to value different interpretations in this course.
20. I have developed skills needed in this field.

Fall 2007 Open-Ended Questions

1. What were the best aspects of this course?

2. What changes could be made to improve the course?

3. Describe the aspects of the teacher's teaching that were most effective.

4. Describe the aspects of the teacher's teaching that could be improved.

Appendix B

Validity Questions

Fall 2007 Validity Questions

1) Your attempted hours this term

2) Your expected grade in the course (A; B; C; D; F; Withdrawal; Pass; Fail)

3) Your reason for taking the course (Major requirement; Major elective; Liberal studies requirement; Minor/related field; Personal interest only)

4) Your level of interest in the subject prior to this course (Very low; Low; Medium; High; Very high)

5) Your overall grade point average at WCU (<2.0; Between 2.0 and 2.4; Between 2.5 and 2.9; Between 3.0 and 3.4; Between 3.5 and 3.7; Above 3.7)

6) The course difficulty relative to other courses is (Very easy; Easy; Medium difficulty; Hard; Very hard)

7) Relative to other courses, the amount of reading in this course is (Very light; Light; Medium; Heavy; Very heavy; Not Applicable)

8) Relative to other courses, the amount of writing in this course is (Very light; Light; Medium; Heavy; Very heavy; Not Applicable)

9) Relative to other courses, the overall work load in this course is (Very light; Light; Medium; Heavy; Very heavy; Not Applicable)

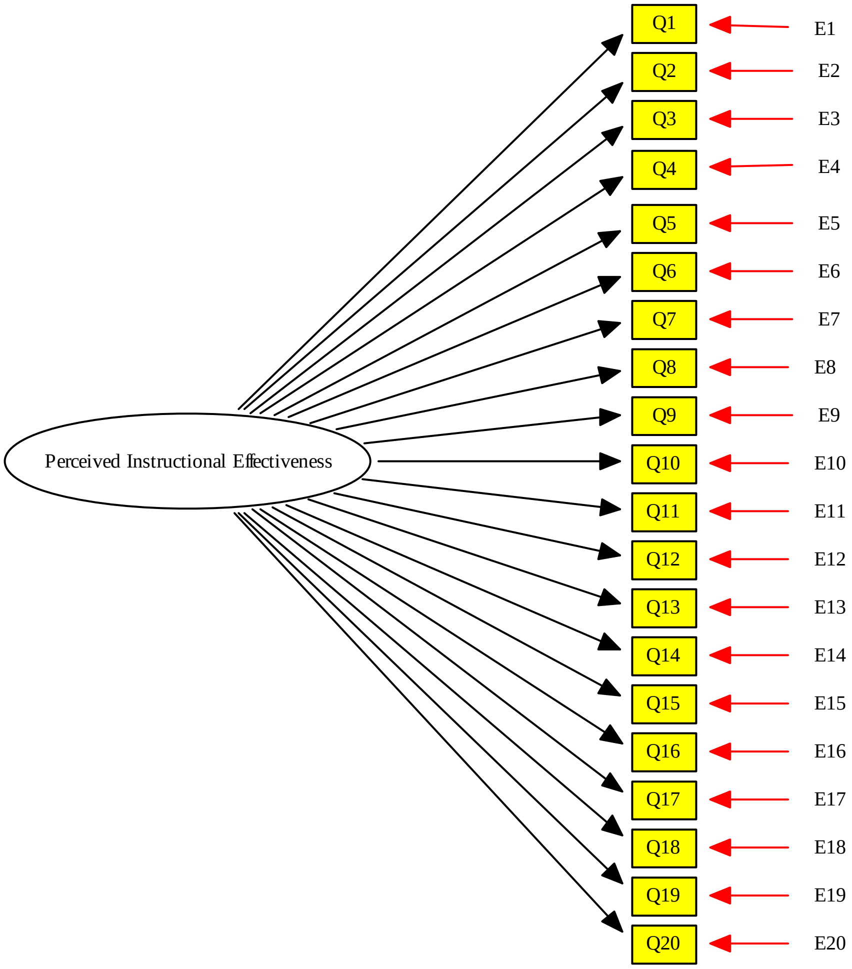
10) Relative to other courses the pace of this course is (Too slow; Slow; About right; Fast; Too fast; Not Applicable)

11) How many hours of work per week are required outside of class (0 to 2 hours per week; 3 to 5 hours per week; 6 to 8 hours per week; 9 to 11 hours per week; over 11 hours per week; Not Applicable)

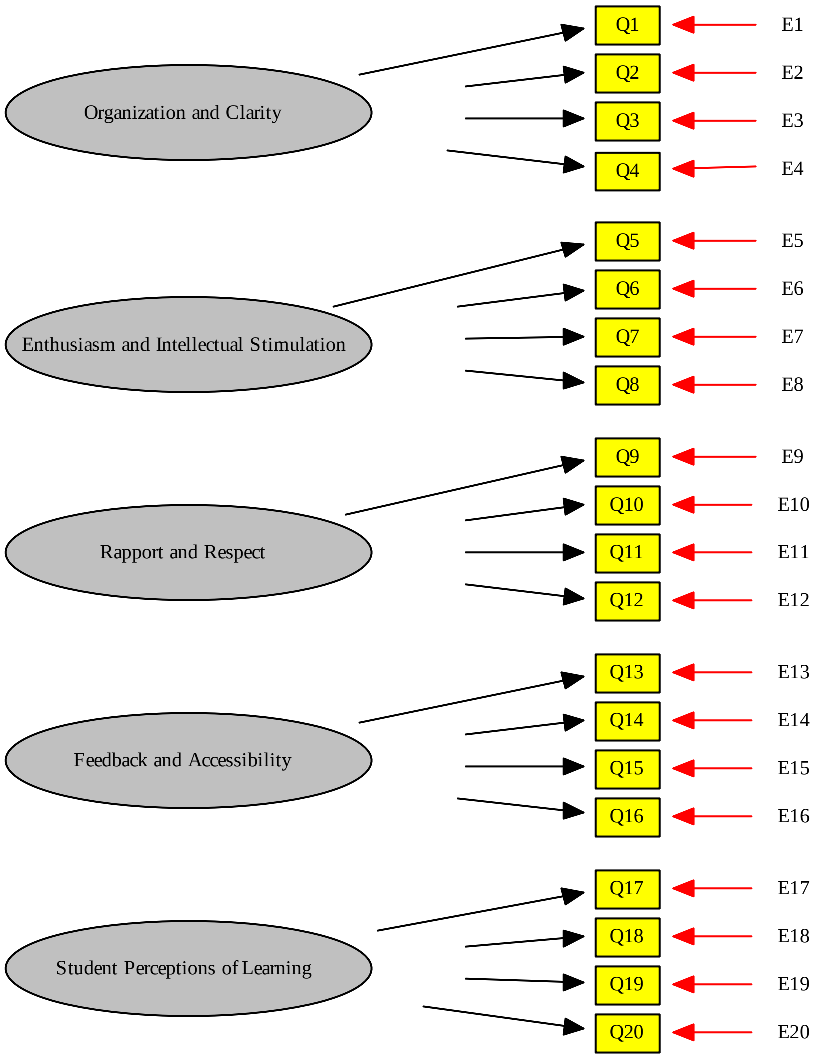
Appendix C

CFA Model Diagrams

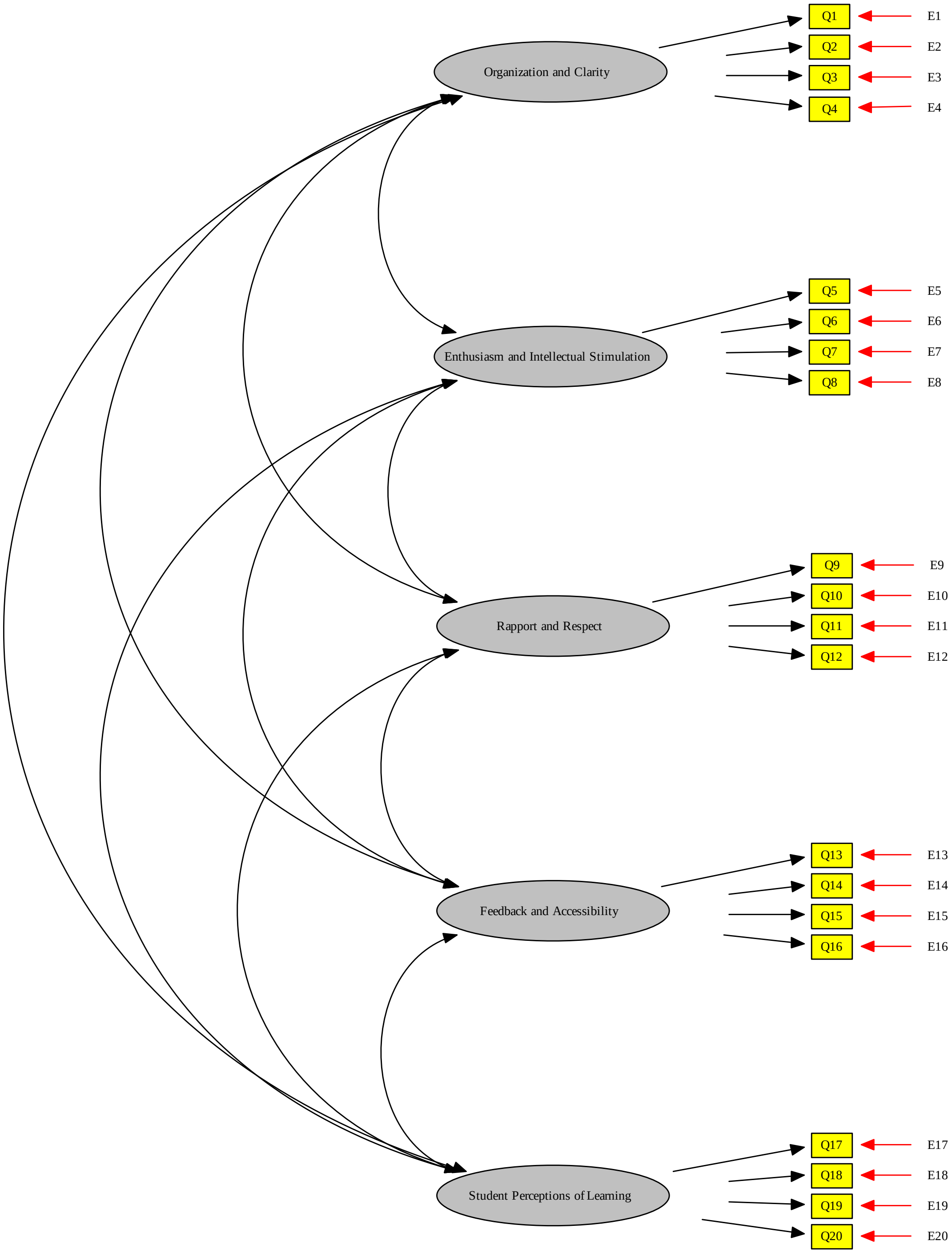
One-Factor Model



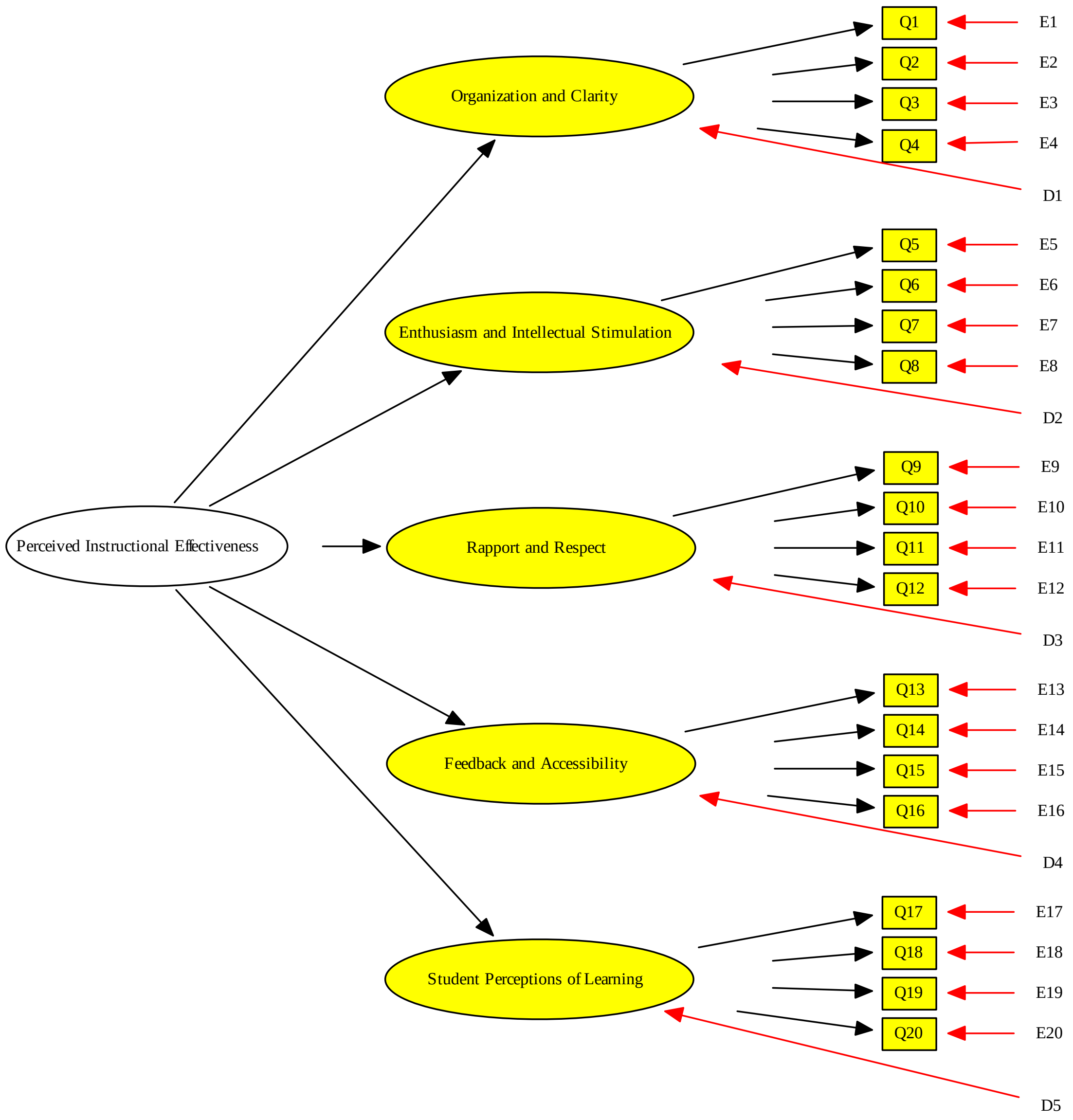
Five-Factor Orthogonal Model



Five-Factor Oblique Model



Second-Order Model



Appendix D

Descriptive Statistics

Table D1

Activities Course Form Descriptive Statistics

	Fall 2007			Spring 2008		
Question #	N	Mean	Stdev	N	Mean	Stdev
1	478	4.57	0.69	126	3.51	0.76
2	521	4.37	0.88	126	3.60	0.68
3	504	4.40	0.88	126	3.53	0.77
4	519	4.35	0.87	125	3.60	0.68
5	482	4.50	0.78	124	3.52	0.78
6	511	4.27	0.95	125	3.52	0.78
7	504	4.43	0.81	125	3.54	0.75
8	464	1.58	0.78	124	3.55	0.76
9	475	4.25	0.96	126	3.49	0.78
10	483	4.20	1.00	126	3.42	0.82
11	486	4.40	0.87	125	3.54	0.74
12	462	4.16	1.08	126	3.45	0.77
13	492	4.42	0.79	125	3.62	0.62
14	496	4.40	0.83	125	3.59	0.65
15	487	4.30	0.92	125	3.59	0.61
16	504	4.35	0.92	126	3.51	0.80
17	494	4.43	0.79	126	3.50	0.79
18	494	4.42	0.82	126	3.49	0.79
19	495	4.38	0.85	126	3.52	0.77
20	493	4.39	0.88	125	3.54	0.75

Table D2

Independent Research Course Form Descriptive Statistics

	Spring 2007			Fall 2007			Spring 2008		
Question #	N	Mean	Stdev	N	Mean	Stdev	N	Mean	Stdev
1	25	4.44	0.92	93	4.65	0.64	96	3.73	0.49
2	25	4.44	0.87	92	4.70	0.61	96	3.75	0.50
3	25	4.40	0.87	90	4.63	0.73	96	3.66	0.65
4	25	1.28	0.68	91	4.52	0.81	96	3.66	0.54
5	25	4.64	0.86	92	4.73	0.58	96	3.85	0.35
6	25	4.52	0.87	93	4.61	0.74	96	3.80	0.47
7	25	4.56	0.87	93	4.70	0.64	94	3.77	0.45
8	25	4.84	0.37	93	4.52	0.84	95	3.54	0.71
9	25	4.60	0.87	92	4.76	0.58	96	3.88	0.33
10	25	4.60	0.87	93	4.76	0.58	95	3.86	0.38
11	25	4.72	0.54	92	4.78	0.46	96	3.86	0.37
12	25	4.80	0.41	93	4.78	0.46	95	3.84	0.42
13	25	4.68	0.75	93	4.67	0.66	96	3.72	0.50
14	25	4.60	0.87	90	4.52	0.75	96	3.70	0.58
15	25	4.72	0.74	93	4.76	0.48	96	3.78	0.42
16	25	4.60	0.87	93	4.68	0.59	96	3.75	0.54
17	25	4.68	0.85	87	4.72	0.52	94	3.69	0.59
18	25	4.64	0.86	91	4.70	0.55	96	3.72	0.50
19	25	4.60	0.87	89	4.62	0.63	96	3.61	0.67
20	24	4.54	0.88	91	4.66	0.58	95	3.72	0.48

Table D3

Internship, Practica, and Clinical Course Form Descriptive Statistics

	Spring 2007			Fall 2007			Spring 2008		
Question #	N	Mean	Stdev	N	Mean	Stdev	N	Mean	Stdev
1	7	4.86	0.38	277	4.44	0.86	281	3.52	0.63
2	5	4.80	0.45	263	4.44	0.84	279	3.50	0.65
3	5	4.60	0.55	273	4.45	0.78	277	3.50	0.67
4	7	4.71	0.49	281	4.54	0.75	281	3.53	0.65
5	6	4.50	0.55	279	4.57	0.77	280	3.56	0.64
6	6	4.83	0.41	279	4.57	0.76	280	3.51	0.68
7	5	3.80	1.64	278	4.50	0.81	279	3.43	0.74
8	6	2.00	1.10	275	4.34	0.92	277	3.38	0.74
9	6	4.50	0.55	275	4.52	0.76	281	3.47	0.68
10	7	4.29	0.49	274	4.51	0.78	281	3.44	0.71
11	5	4.40	0.55	278	4.50	0.77	279	3.45	0.69
12	6	3.67	1.37	260	4.56	0.72	277	3.58	0.61
13	7	4.14	0.69	271	4.50	0.75	279	3.52	0.62
14	7	3.86	0.90	269	4.52	0.75	278	3.49	0.66
15	6	4.17	0.75	276	4.51	0.80	277	3.50	0.68
16	3	4.67	0.58	253	4.42	0.81	277	3.47	0.67
17	6	4.17	0.75	276	4.51	0.80	275	3.45	0.69
18	7	4.14	0.69	275	4.49	0.74	274	3.45	0.68
19	7	5.00	0.00	280	4.59	0.75	279	3.56	0.64
20	6	4.83	0.41	279	4.57	0.75	277	3.56	0.64

Table D4
Laboratory Course Form Descriptive Statistics

	Spring 2007			Fall 2007			Spring 2008		
Question #	N	Mean	Stdev	N	Mean	Stdev	N	Mean	Stdev
1	376	4.17	0.99	1,109	4.39	0.81	1,154	3.54	0.61
2	376	4.19	0.98	1,114	4.41	0.77	1,153	3.51	0.61
3	375	3.93	1.12	1,105	4.27	0.91	1,153	3.44	0.68
4	375	4.09	1.06	1,103	4.34	0.83	1,149	3.49	0.65
5	376	3.78	1.09	1,105	4.08	0.92	1,152	3.33	0.70
6	376	3.97	1.12	1,109	4.34	0.85	1,151	3.51	0.63
7	376	3.93	1.11	1,105	4.28	0.86	1,151	3.45	0.67
8	371	3.84	1.12	1,097	4.24	0.92	1,154	3.41	0.73
9	373	4.49	0.79	1,071	4.56	0.64	1,145	3.59	0.54
10	367	4.05	1.05	1,082	4.21	0.97	1,144	3.38	0.74
11	376	4.29	0.88	1,109	4.48	0.73	1,150	3.55	0.61
12	363	4.03	1.02	1,077	4.28	0.85	1,146	3.45	0.64
13	375	3.90	1.24	1,106	4.33	0.80	1,153	3.43	0.68
14	374	3.89	1.16	1,106	4.22	0.90	1,151	3.38	0.70
15	375	4.14	0.95	1,104	4.35	0.78	1,149	3.46	0.63
16	373	3.84	1.17	1,101	4.23	0.88	1,150	3.39	0.71
17	376	4.03	1.04	1,107	4.31	0.84	1,148	3.45	0.65
18	376	3.71	1.21	1,108	4.11	0.96	1,153	3.33	0.75
19	365	3.96	1.08	1,089	4.21	0.88	1,150	3.40	0.66
20	375	4.02	1.04	1,108	4.28	0.85	1,149	3.45	0.66

Table D5

Online Course Form Descriptive Statistics

	Spring 2007			Fall 2007			Spring 2008		
Question #	N	Mean	Stdev	N	Mean	Stdev	N	Mean	Stdev
1	196	4.16	1.05	1,507	4.03	1.18	837	3.44	0.70
2	196	4.21	1.09	1,495	4.15	1.08	838	3.48	0.71
3	194	4.15	1.02	1,471	4.12	1.07	836	3.51	0.62
4	177	4.04	1.13	1,416	4.28	1.06	830	3.57	0.61
5	195	4.00	1.08	1,417	4.07	1.13	833	3.46	0.67
6	195	3.91	1.13	1,413	4.01	1.15	837	3.43	0.70
7	193	4.16	1.08	1,351	4.18	1.15	838	3.55	0.62
8	196	3.83	1.22	1,397	3.92	1.19	830	3.36	0.78
9	187	4.19	0.97	1,329	4.24	1.03	840	3.54	0.57
10	189	4.21	0.99	1,359	4.27	1.02	836	3.57	0.58
11	193	4.00	1.11	1,363	4.14	1.05	838	3.49	0.62
12	184	3.98	1.09	1,332	4.08	1.08	837	3.44	0.70
13	185	3.99	1.16	1,389	3.89	1.20	834	3.25	0.90
14	183	4.02	1.11	1,381	4.02	1.16	832	3.39	0.75
15	186	4.02	1.09	1,423	4.14	1.02	827	3.47	0.68
16	186	3.89	1.19	1,420	4.00	1.10	829	3.40	0.73
17	195	4.07	1.02	1,395	4.10	1.07	830	3.45	0.66
18	193	3.97	1.13	1,338	4.17	1.06	837	3.49	0.64
19	193	3.84	1.15	1,373	4.01	1.10	836	3.40	0.71
20	195	3.95	1.08	1,373	4.07	1.07	836	3.43	0.67

Table D6
Seminar Course Form Descriptive Statistics

	Spring 2007			Fall 2007			Spring 2008		
Question #	N	Mean	Stdev	N	Mean	Stdev	N	Mean	Stdev
1	35	4.63	0.55	1,723	4.52	0.73	237	3.67	0.52
2	35	4.63	0.55	1,727	4.41	0.81	237	3.65	0.56
3	34	4.56	0.79	1,714	4.48	0.81	235	3.60	0.65
4	34	4.29	0.76	1,716	4.25	0.91	236	3.46	0.74
5	34	4.38	0.92	1,719	4.25	0.93	234	3.44	0.72
6	35	4.43	0.95	1,715	4.21	0.97	233	3.39	0.78
7	34	4.65	0.73	1,724	4.55	0.74	235	3.69	0.53
8	34	4.41	0.78	1,720	4.35	0.89	234	3.56	0.67
9	35	4.66	0.48	1,703	4.28	0.87	236	3.56	0.68
10	33	4.58	0.61	1,679	4.09	1.06	236	3.56	0.67
11	34	4.79	0.59	1,726	4.54	0.70	237	3.69	0.50
12	35	4.77	0.43	1,722	4.48	0.78	236	3.63	0.58
13	35	4.49	0.82	1,716	4.43	0.74	236	3.61	0.64
14	35	4.57	0.56	1,711	4.36	0.84	234	3.67	0.53
15	34	4.35	0.85	1,711	4.26	0.91	233	3.54	0.69
16	35	4.60	0.69	1,714	4.39	0.84	234	3.58	0.63
17	34	4.53	0.75	1,721	4.38	0.81	234	3.52	0.67
18	33	4.45	0.75	1,680	4.13	0.93	234	3.46	0.69
19	35	4.46	0.74	1,723	4.30	0.83	234	3.53	0.64
20	33	4.52	0.67	1,715	4.37	0.79	236	3.59	0.59

Table D7
Standard Lecture Course Form Descriptive Statistics

	Spring 2007			Fall 2007			Spring 2008		
Question #	N	Mean	Stdev	N	Mean	Stdev	N	Mean	Stdev
1	1,802	4.43	0.90	10,782	4.44	0.83	9,118	3.56	0.62
2	1,805	4.13	1.11	10,821	4.19	1.04	9,119	3.39	0.77
3	1,809	4.23	1.02	10,832	4.28	0.95	9,110	3.45	0.72
4	1,807	4.29	1.01	10,847	4.30	0.95	9,115	3.47	0.70
5	1,809	4.44	0.90	10,807	4.49	0.80	9,097	3.60	0.62
6	1,804	4.08	1.16	10,788	4.10	1.10	9,102	3.33	0.81
7	1,807	4.12	1.12	10,810	4.16	1.04	9,095	3.37	0.78
8	1,804	4.08	1.12	10,801	4.15	1.03	9,106	3.38	0.77
9	1,804	4.23	0.99	10,716	4.27	0.93	9,097	3.44	0.71
10	1,776	4.22	0.96	10,625	4.29	0.90	9,082	3.44	0.69
11	1,778	4.15	1.04	10,506	4.21	0.99	9,068	3.40	0.75
12	1,790	4.30	0.93	10,669	4.32	0.90	9,109	3.48	0.68
13	1,800	4.09	1.09	10,688	4.17	0.99	9,098	3.34	0.77
14	1,806	4.27	0.94	10,769	4.29	0.91	9,092	3.44	0.70
15	1,806	4.21	1.02	10,783	4.28	0.93	9,097	3.43	0.72
16	1,801	4.00	1.17	10,760	4.11	1.06	9,097	3.32	0.81
17	1,809	4.23	1.02	10,761	4.29	0.94	9,101	3.44	0.72
18	1,807	4.21	1.00	10,724	4.27	0.93	9,089	3.43	0.71
19	1,806	4.17	1.02	10,709	4.22	0.95	9,079	3.40	0.73
20	1,796	4.22	0.98	10,606	4.27	0.91	9,084	3.45	0.69

Table D8
Studio-Performance Course Form Descriptive Statistics

	Fall 2007			Spring 2008		
Question #	N	Mean	Stdev	N	Mean	Stdev
1	430	4.55	0.78	538	3.87	0.34
2	426	4.38	1.01	530	3.80	0.44
3	425	4.36	1.03	538	3.79	0.48
4	414	4.35	1.03	535	3.80	0.47
5	432	4.57	0.78	539	3.80	0.52
6	431	4.51	0.89	540	3.73	0.61
7	431	4.52	0.84	540	3.76	0.60
8	428	4.53	0.86	540	3.81	0.53
9	423	4.38	1.01	540	3.75	0.54
10	419	4.32	0.99	540	3.72	0.54
11	421	4.23	1.13	540	3.61	0.72
12	420	4.21	1.05	540	3.61	0.67
13	410	4.31	1.02	532	3.70	0.63
14	409	4.44	0.87	532	3.77	0.50
15	417	4.44	0.92	537	3.73	0.61
16	419	4.38	1.02	533	3.78	0.55
17	425	4.55	0.80	534	3.72	0.57
18	427	4.54	0.79	528	3.76	0.55
19	418	4.48	0.85	537	3.77	0.52
20	426	4.58	0.71	537	3.75	0.56