Utah State University

DigitalCommons@USU

All Graduate Theses and Dissertations

Graduate Studies

12-2010

Psychometric Properties of Postsecondary Students' Course Evaluations

Michael J. Drysdale
Utah State University

Follow this and additional works at: <https://digitalcommons.usu.edu/etd>

Part of the [Educational Assessment, Evaluation, and Research Commons](#), [Educational Psychology Commons](#), and the [Statistics and Probability Commons](#)

Recommended Citation

Drysdale, Michael J., "Psychometric Properties of Postsecondary Students' Course Evaluations" (2010).
All Graduate Theses and Dissertations. 783.
<https://digitalcommons.usu.edu/etd/783>

This Dissertation is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Theses and Dissertations by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.

PSYCHOMETRIC PROPERTIES OF POSTSECONDARY
STUDENTS' COURSE EVALUATIONS

by

Michael J. Drysdale

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Psychology

Approved:

Karl R. White, Ph.D.
Major Professor

J. Nicholls Eastmond, Ph.D.
Committee Member

James T. Dorward, Ph.D.
Committee Member

Kerry E. Jordan, Ph.D.
Committee Member

Mark S. Innocenti, Ph.D.
Committee Member

Byron R. Burnham, Ed.D.
Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2010

Copyright © Michael J. Drysdale 2010

All Rights Reserved

ABSTRACT

Psychometric Properties of Postsecondary
Students' Course Evaluations

by

Michael J. Drysdale, Doctor of Philosophy

Utah State University, 2010

Major Professor: Karl R. White, PhD
Department: Psychology

Several experts in the area of postsecondary student evaluations of courses have concluded that they are stable or reliable measures as well as being measures that provide ways of making valid inferences regarding teacher effectiveness. Often these experts have offered these conclusions without supporting evidence. Surprisingly, a thorough review of the literature revealed very few reported test-retest reliability studies of course evaluations and the results from these studies are contradictory. In the area of validity, the conclusions offered by scholars who conducted meta-analyses of multisection course studies are inconsistent. This leads to the following two research questions:

1. What is the test-retest reliability over a 3-week period of the course evaluation currently employed at Utah State University?
2. Can results of the course evaluation employed at Utah State University be used to make valid inferences about a teacher's effectiveness?

Two parts of a study were conducted to answer these questions. First, a test-retest reliability part was conducted with students from courses at Utah State University, employing a 3-week time lapse between administrations of the course evaluations. Second, a multisection course validity part was conducted using existing student ratings data and final examination scores for 100 sections of MATH 1010 over a 5-year period. Correlational analyses were conducted on the resulting data from both studies. Test-retest reliability coefficients ranging from 0.64 to 0.94 were found. In the second study, the correlation coefficients from the validity study ranged from -0.39 to 0.71, with a mean coefficient of 0.14 and 0.11 for final examination score by instructor rating and final examination score by course rating, respectively. Results from both parts of the study suggest that the course evaluation used at USU is not reliable and that results of the course evaluation do not provide information that can be used to make valid inferences regarding teacher effectiveness.

(121 pages)

ACKNOWLEDGMENTS

First and foremost, I want to thank my wife for her love and support. She has been extremely patient and helpful throughout this entire process. Also, I want to thank my children, McKenna, Hannah, Dallin, Eliza, and Esther, who have also been patient as I have not spent as much time with them as I, or they, would have liked. I love you all!

I am also very thankful for the guidance and support that Dr. Karl White has given me. Without his help I know that this project never would have been completed.

Lastly, I wanted to thank the AAA Office and the Department of Mathematics at Utah State University. Once again, without their help this study would never have come to fruition.

Thank you!

Michael J. Drysdale

CONTENTS

vi

	Page
ABSTRACT.....	iii
ACKNOWLEDGMENTS.....	v
LIST OF TABLES.....	viii
LIST OF FIGURES.....	x
CHAPTER	
I. INTRODUCTION.....	1
II. LITERATURE REVIEW.....	5
Ubiquity of Course Evaluations.....	5
Researching Course Evaluations.....	8
Reliability.....	10
Validity.....	23
Overall Course Rating with Student Achievement.....	26
Overall Instructor Rating with Student Achievement.....	27
Utah State University Course Evaluation.....	40
Conclusions.....	41
III. METHOD.....	43
Test-Retest Reliability Design.....	44
Multisection Course Design.....	45
IV. RESULTS AND DISCUSSION.....	49
Test-Retest Reliability Design.....	49
Multisection Course Design.....	63
V. CONCLUSIONS.....	67
Expert Opinion.....	67
Response Rate.....	68
Test-Retest Reliability.....	69
Validity.....	70
Limitations.....	71

Implications.	74
REFERENCES.	76
APPENDICES.	81
Appendix A: List of Institutional Research Centers.	82
Appendix B: Email to Offices of Institutional Research.	84
Appendix C: Follow-up Email to Offices of Institutional Research. .	86
Appendix D: Table of Studies Included (and those not included) in Cohen (1981) and Clayson (2009).	88
Appendix E: USU Course Evaluation.	90
Appendix F: Letter of Information.	92
Appendix G: Final Exam Means and Instructor/Course Ratings Separated by Instructor.	95
Appendix H: Scatterplots for Final Exam Mean x Course Rating and Final Exam Mean x Instructor Rating.	97
CURRICULUM VITAE.	108

LIST OF TABLES

Table		Page
1	Quotes from Prominent Researchers About the Psychometric Properties of Course Evaluations	9
2	Overall and Marsh (1980) Mean Differences Between Student Ratings. . . .	15
3	Median Correlation Coefficients for Instructor Ratings at Times 1, 2, and 3 from West (1988)	17
4	Comparison of Test-Retest Reliability Studies of Course Evaluations.	19
5	Comparison of Studies Included in the Cohen (1981) and Clayson (2009) Meta-Analyses.	32
6	Data from Hypothetical Data Sets 1 and 2 Looking at Validity of Course Evaluations.	34
7	Comparison of Studies Included in Cohen (1981) and Clayson (2009) That Included Means (ratings/final exams)	36
8	Means Reported in those Studies Included in Cohen (1981) and Clayson (2009)	37
9	Results from Cronbach's Alpha Analysis of the Items from the Utah State University Course Evaluation.	41
10	Number of Sections of MATH 1010 and the Available Data from the Last Five Years.	48
11	Sample Sizes from the Course Sections Included in the Test-Retest Reliability Study.	50
12	Mean Ratings and Standard Deviations from the Course Sections Included in the Test-Retest Reliability Study.	50
13	Course Evaluation Completion Data from Spring 2010 Semester at Utah State University.	53
14	Test-Retest Reliability Correlation Coefficients.	54

Table		ix Page
15	Descriptive Statistics for Retest Reliability Coefficients for Ten Test Types Taken from Charter (2003)	56
16	Comparison of Cronbach's Alpha Using Data from the Present Study and from USU (2010).	57
17	Retaining and Terminating Instructors Based on a Reliability Coefficient of 0.90.	58
18	Retaining and Terminating Instructors Based on a Reliability Coefficient of 0.70.	59
19	Hypothetical Example of Using Course Evaluations with a Reliability Coefficient of 0.74 to Terminate Instructors at Utah State University.	60
20	Hypothetical Example of Using Course Evaluations with a Reliability Coefficient of 0.78 to Terminate Instructors at Utah State University.	61
21	Comparison of Reliability Coefficients for Overall- and Mean-Course and Instructor Ratings.	62
22	Multisection Course Validity Study Descriptive Statistics and Correlation Coefficients.	64
23	Differences in Data from Students Who Included Their ID Number and Those Who Did Not.	73

LIST OF FIGURES

Figure		Page
1	Distribution of course/achievement correlations for 22 courses taken from Cohen, 1981.	28
2	Distribution of instructor/achievement correlations for 67 courses taken from Cohen, 1981.	28
3	Distribution of student rating/achievement correlations for 42 courses based on data presented by Clayson (2009)	31

CHAPTER I

INTRODUCTION

Course evaluations completed by students, or student ratings as they are commonly known, are an important, ubiquitous part of teacher evaluation in the current higher education scene in America. Most people agree that the use of student ratings has been increasing over the past 25 years (Lin, McKeachie, & Tucker, 1984; Penny, 2003) and that most colleges and universities across the United States use the results of course evaluations in judging the effectiveness of their teachers (Benassi & Seidel, 2006; Cashin, 1989; McKeachie & Svinicki, 2005). Most experts in this area have also concluded that course evaluations are, (a) the most psychometrically sound instrument currently available for measuring teacher effectiveness (McKeachie, 1997; Scriven, 1988), (b) reliable measures of teacher effectiveness (Overall & Marsh, 1980; Witt & Burdalski, 2003), and (c) a reasonable means for making valid inferences regarding teacher effectiveness (e.g., Cohen, 1981; Feldman, 1989).

Two problems exist in the literature that has been used to draw these conclusions. First, only four test-retest studies of the reliability of course evaluations could be found (Overall & Marsh, 1980; Scott, Hunt, & Greig, 1986; West, 1988; Witt & Burdalski, 2003). Methodological problems in these studies (i.e., long time lapse between the two testing events and/or only reporting differences in means instead of correlations between results at the two testing times), along with contradictory results and conclusions, do not provide convincing evidence for or against the reliability of course evaluations.

Second, validity issues related to course evaluations have not been adequately addressed. A multisection course validity design has been used most frequently to

evaluate the validity of inferences about teacher effectiveness based on course evaluations. In a multisection course validity design, the same course is taught by multiple instructors during the same semester and students in the various sections of the course all take the same final examination. Then, the ratings from the course evaluations are correlated with student scores on the final exam. Positive correlations are viewed as evidence that teachers who are rated more positively by students are, in fact, more effective teachers because the students in those courses have higher average scores on the final exam.

The most widely cited work in this area is Cohen's (1981) meta-analysis of multisection studies, in which he analyzed results from 68 studies reported in 41 articles. Cohen found that correlations between ratings and student exam scores ranged from -0.30 to +0.90, with a mean of +0.43. Based on these data Cohen (1981) concluded that course evaluations could be used to make valid inferences regarding teacher effectiveness. In his words, results from course evaluations provided "strong support for the validity of student ratings as measures of teaching effectiveness" (p. 300). Additionally, in 2009, Clayson conducted a meta-analysis of multisection course validity studies, similar to the study conducted by Cohen (1981), including new studies and studies from other disciplines that had not been included in Cohen's (1981) meta-analysis. Clayson (2009), who included a total of 42 studies reported in 17 articles, found mostly positive relationships between final exam scores and ratings on the course evaluations, including 13 of the 42 that were statistically significant. Correlation coefficients ranged from -0.75 to +0.91, with an average of +0.33. Clayson (2009) concluded that even though there was

a positive relation between learning and ratings, “the average association’s magnitude is small ($r = 0.33$) and inconsistent” (p. 23).

A problem with both of these analyses was that no means were reported (for ratings or for grades). This is an issue because it is possible to have high correlation coefficients between student ratings and student exam scores and have very little difference in the exam scores or the ratings between the course sections. In such a case, student evaluations would not clearly discriminate between effective and less effective instructors. All but four of the studies included in Cohen’s (1981) and Clayson’s (2009) meta-analyses were obtained. Of the 48 studies obtained, only three reported means, but did not analyze them in any manner. The inclusion and analysis of the differences in mean ratings and mean exam scores would provide more evidence about whether the results of students’ course evaluations can be used to distinguish between effective and less effective instructors.

To address these two issues the present study consisted of three parts. First, the offices of institutional research at 50 universities in the United States were contacted in an effort to discover unpublished research about the psychometric properties of student evaluations. Secondly, because reliability of measurement is a necessary precursor to validity, test-retest correlations of course evaluations over three time periods of about three weeks each for multiple classes were examined. In the final part of this study existing data were used to conduct a multisection course validity study. This part of the study included and analyzed mean ratings and mean exam scores, in addition to reporting correlation coefficients.

Results from this study provide a better understanding of the psychometric properties of the course evaluation currently employed at Utah State University (USU) with important implications for similar course evaluations employed at other universities across the United States.

CHAPTER II

LITERATURE REVIEW

This chapter will discuss the ubiquity of course evaluations, the expert opinions about course evaluations, as well as the empirical research that has been conducted looking at the reliability and validity of course evaluations.

Ubiquity of Course Evaluations

Across the country, near the end of every semester hundreds of thousands of college students complete course evaluations. For any one class, this exercise usually requires 10 minutes or less for each student, but it happens multiple times for each student each semester. For example, at USU, with an on-campus enrollment of approximately 16,000 students, more than 50,000 course evaluations are completed each semester.

Much time, energy, and money is invested in collecting, analyzing, and reporting these course evaluations. This process is in place for a few reasons, including promotion and tenure decisions about individual faculty members, helping instructors improve their teaching, and providing students with information about instructors when they register for courses. Course evaluations, or the student ratings that come from them, have become an important part in this process of evaluating instructors. Many experts have drawn conclusions about the use, reliability, and validity of course evaluations, but have almost always done so without providing concrete evidence. For example, both Centra (1993) and Penny (2003) stated that course evaluations were introduced in the 1920s, but

provided no data or sources to back up that assertion. Furthermore, Lin and colleagues (1984) noted that the use of course evaluations has been steadily increasing ever since their introduction, with a marked increase during the reforms in higher education brought about by student protests of the late 1960s. Again, this is done without any supporting evidence.

Cashin (1989) pointed out that almost every university states that teaching is its primary purpose and, therefore, every university evaluates its instructors in one form or another. He further claimed that universities have “increasingly” incorporated the use of course evaluations into these evaluation efforts. Unfortunately, he provided no data to support those claims. Centra (1993) stated that “rare is the American college or university that does not currently use student evaluations of teaching in one way or another” (p. 47). He also noted that this trend had steadily increased for the past 25 years. Again, these claims are made without reference to any specific data or research. Additionally, Benassi and Seidel (2006) referred to the increased use of course evaluations that had occurred over the past 25 years or so. They stated that this increase in use had developed to the point of ubiquity, but provided no supporting evidence. Kolitch and Dean (1999) stated that course evaluations “are viewed as the ‘de facto gold standard of retention, tenure, promotion decisions,’ giving them a taken-for-granted, common-sense, institutional legitimacy” (p. 27). Furthermore, McKeachie and Svinicki (2005) claimed that course evaluations “are now administered in almost all colleges and universities in the United States and are becoming common in other countries” (p. 326). Again, no data were offered in support of these statements. This could be the case because it is everyone’s experience that course evaluations are used, and therefore, the scholars do not recognize a

need for providing support for the claims they make. In contrast, it could be that no empirical evidence has been collected or published.

In an effort to find supporting evidence for these claims a search of both the ERIC and PsychInfo databases was conducted. This search was done as follows: the key words *student ratings and course evaluations* were input into one field, while *frequency, use,* and *ubiquity* were input into the second field. A few studies were located and are discussed next.

There are approximately 4,000 postsecondary institutions in the United States. In 1975, Bejar surveyed 333 of these institutions about their use of student ratings of instructors. Of the 228 responding schools, 168 (73.7%) were using or planned to use student ratings of instructors in the near future.

More recently, Seldin (1993) summarized three studies he had conducted looking at teacher evaluation practices, as they related to promotion and tenure, salary increase, and retention, at liberal arts colleges ($n = 600$) across the United States. The studies were conducted in 1973, 1983, and 1993. Relevant to the present discussion was the use of the student ratings that came from the course evaluations. The percent of colleges who reported using course evaluations went from 29% in 1973, to 68% in 1983, to 86% in 1993. In his 1984 report, Seldin stated that “student ratings are now ‘always used’ in personnel decisions by more than two thirds of the colleges” (p. 32). Surveys conducted by Seldin (1975, 1984, 1993) show a steady increase (29% in 1973 to 86% in 1993) in the use of course evaluations at liberal arts colleges during the 1970s, 1980s, and 1990s.

Because Seldin (1993) was the most recent report found via *ERIC* and *PsychInfo* searches, *Google Scholar* was also employed. Utilizing the “cited by” link in *Google*

Scholar, it was found that 91 articles had cited the Seldin (1993) report. The titles and abstracts of the resulting 91 articles were reviewed but none were found that had conducted research looking at the use of course evaluations. Most common were opinion articles without the inclusion of conducted research. Additionally, other studies looking at instructors, their views on student perspectives or other classroom specific behaviors were found.

Some support was found for the conclusions given by “experts” regarding course evaluations and their implementation. This evidence, even though 17 years old or more, does suggest that there was an increased use of course evaluations from the 1970s to the 1990s.

Researching Course Evaluations

Because results of course evaluations are used to make important decisions, it is important to understand their psychometric properties. For example, the value of using course evaluations in making decisions about the promotion or tenure of a particular professor is linked to the reliability of results from these evaluations. Based on reviews of research related to the psychometric properties of course evaluations (e.g., Cashin, 1995; Marsh, 2007) most readers would probably conclude that the course evaluations currently employed at colleges and universities across the United States provide a psychometrically sound way of evaluating college professors and instructors. As noted by the quotes in Table 1, many scholars have concluded that course evaluations are the most valid instrument currently available to measure teacher effectiveness.

Table 1

Quotes from Prominent Researchers About the Psychometric Properties of Course Evaluations

Citation	Quote	Page number
Cashin, 1988, 1995	“...student ratings tend to be statistically reliable, valid, and relatively free from bias or the need for control; probably more so than any other data used for evaluation.”	6
Marsh, 2007	Cites and agrees with Cashin (1988): “...student ratings tend to be statistically reliable, valid, and relatively free from bias, probably more so than any other data used for faculty evaluation.”	343-344
McKeachie, 1997	“...student ratings are the single most valid source of data on teaching effectiveness.”	1219
Scriven, 1988	“...student ratings are not only a valid, but often the only valid, way to get much of the information needed for most evaluations—and for many useful descriptions—of teaching.”	14-15

Both McKeachie (1997) and Scriven (1988) provide no supporting evidence or data when offering their conclusions. Both Cashin (1995) and Marsh (2007) make extensive use of the Cohen (1981) meta-analysis in their sections on validity. More about the Cohen (1981) meta-analysis is included in the validity section of the present literature review.

To provide an understanding of the research underlying the conclusions drawn by respected scholars in this area, studies of the psychometric properties of course evaluations will be summarized, with a particular emphasis on test-retest reliability studies and multisection course validity studies.

Reliability is a necessary but not sufficient precursor to validity. Therefore, the reliability of course evaluations will be discussed first, followed by a discussion of validity issues related to course evaluations.

Reliability

Reliability is a measure of how consistent an instrument is over times or settings or people. For example, results of course evaluations (or student ratings of teacher effectiveness), would be deemed to have high reliability if, when the same student is evaluating the same instructor at two different times, the ratings were the same or very similar.

This concept is commonly referred to as test-retest reliability. When applied to the case of student evaluations, it would be assessed by asking students to rate the instructor of a course in which they are currently enrolled. Later, after a designated amount of time has elapsed, the same students are asked to rate the same instructor again. The data from the two times are used to determine the correlation between the results of the two administrations. High correlation coefficients would demonstrate that the results are consistent, or that the instrument has high reliability.

A significant point in test-retest reliability methods that needs to be raised is that the amount of time that elapses between test and retest is important. If too much time elapses between administrations of the course evaluations, any number of things may occur which would have an impact on the ratings. A couple of examples could include the following: students may take other courses from the same instructor, which would impact their memory and perceptions of the instructor; students may learn more about the

subject and decide that “the instructor really didn’t know what he or she was talking about;” or after completing the first evaluation the student may learn more about effective instructional practices and change his or her mind about how well the instructor teaches. Ideally, test-retest reliability research should be conducted in a way that minimizes the amount of change, in both the instructor’s behavior and the student’s perception of the instructor that occurs between the “tests.” The amount of time that elapses between administrations should be short enough that the conclusion can be made that those kinds of changes have not occurred, but long enough that the students do not remember the responses they gave during the first administration. For student evaluations a period of three weeks should be long enough that students will not remember, but be a short enough period of time for the previously mentioned changes not to occur.

It is better to measure the reliability of students’ course evaluations at the end of the semester. Toward the end of the semester students have 12 to 14 weeks of experience with the instructor. In addition, it is less likely that the instructor will be changing in-class behaviors. Things tend to be more set and less likely to change.

The most oft cited study that used a test-retest methodology with course evaluations was conducted in 1980 by Overall and Marsh, entitled “Students’ Evaluations of Instruction: A Longitudinal Study of Their Stability.” Based on *Scopus* and *Google Scholar* searches it is the most cited study of the reliability of course evaluations, both by those who conducted reliability research and those who have discussed the subject. Searches were conducted in an effort to find studies using similar methodology. Because this study is cited so frequently, a search was also made to find those studies that cited the Overall and Marsh study. *Web of Science* was first considered as the database for

conducting this search, but it only includes articles dating back to 1982. Therefore, the *Scopus* search engine was used. Twelve articles were found that cited the Overall and Marsh study. The titles and abstracts from these 12 studies were reviewed. Some of the articles discussed course evaluations, but none of them conducted a study of the reliability of course evaluations.

Google Scholar was also used to find studies that had cited the Overall and Marsh (1980) study. Fifty-nine articles were found. Their titles and abstracts were reviewed and four (Marsh & Overall, 1981; Scott et al., 1986; West, 1988; Witt & Burdalski, 2003) were found that conducted studies using a test-retest methodology. Others that cited Overall and Marsh (1980) reported on developing their own faculty evaluation system, how feedback influences instructors, descriptions of effective instructors, or discussions of research on course evaluations. After further investigation it was found that the Marsh and Overall study actually used the same dataset employed by the Overall and Marsh study with the same test-retest analysis reported in both. Therefore, the Marsh and Overall study will not be reviewed. As a result, the following four published studies were found to have examined test-retest reliability of course evaluations: Overall and Marsh, Scott et al., West, and Witt and Burdalski.

Due to this small number of published test-retest reliability studies, further attempts were made to locate additional, unpublished studies. To that end the directors of institutional research centers from 50 universities from across the United States were contacted (the 10 peer institutions for USU and 40 from the remaining states; these 40 were randomly selected from lists of universities in each state; see Appendix A for a list). An email (see Appendix B) was sent to the director of each center asking whether similar

research had been conducted at the centers and, if so, could access to those reports be gained. A week later a follow-up email (see Appendix C), including a list of the studies already found, was sent to those who had not responded to the first email.

A response was received from 14 of the 50 centers (marked with an asterisk in Appendix A). Most (11 out of the 14) responded by stating that they knew of no such research that had been conducted at their center or institution. Three of the 14 did provide a copy of documents containing research that had been conducted. These documents were reviewed and it was found that none of the documents were reports of test-retest reliability studies of course evaluations like those already found. One report received was found to be of importance. Conducted in 2010 by Dale Pietrzak of the University of South Dakota, the study looked at the test-retest reliability of instructor ratings from two different courses, taught by the same instructor, over a single semester. Results gained from this methodology would help determine the consistency of an instructor from one course to another. This is not the focus of the current research, but can provide further insights into course evaluations and how the results from these measures could be used.

Finally, a third of the institutions ($n = 13$) that did not respond to the emails were randomly selected and all of them contacted by phone (marked with a triangle in Appendix A; making in total 27 that replied or were contacted by phone). After being referred multiple times to different people on each campus, it was found that no one in these 13 additional centers was aware of any similar studies about the reliability or validity of course evaluations that had been done on their campuses. Therefore, a discussion of the four test-retest reliability studies follows.

In 1980, Overall and Marsh employed the first methodology to use the same students evaluating the same course at two different times and, therefore, to consider the test-retest reliability of course evaluations. Prior to 1980, no evidence was found to suggest that researchers had employed this methodology. Overall and Marsh collected data from students in business administration courses. Over the course of the last two years of their studies, students completed end-of-semester course evaluations (consisting of nine items). The items of most importance for the present study were numbers 8: Overall instructor rating, and 9: Overall course rating. These items were rated on a 9-point Likert scale ranging from 1 = *very low* to 9 = *very high*. One year after graduation the same students were contacted and asked to re-evaluate the courses they had taken during their schooling. Therefore, some courses were being re-evaluated three years later, whereas other courses were being evaluated one year later. One hundred sections, with at least eight students in each, were included in the analysis. The overall median correlation coefficients between the ratings on the end-of-semester course evaluation and the re-evaluation (at least one year later) for items 8 and 9 were .84 and .83, respectively. These findings were reported as being statistically significant at the .01 level, showing a high amount of agreement between the ratings. The analysis further included a difference rating between the mean ratings of the end-of-semester course evaluations and the after graduation re-evaluations (see Table 2).

The difference rating for item number 8 was +0.19, whereas the difference rating for item number 9 was -0.15. Both of these findings were reported as being statistically significant at the .05 level, showing that there were differences between the overall rating of instructors and courses at the end of the semester and at least one year later. But, as

Table 2

Overall and Marsh (1980) Mean Differences Between Student Ratings

	End-of-Semester	At least 1 year later	Difference
Item 8			
Overall instructor rating	6.55	6.74	0.19
Item 9			
Overall course rating	6.65	6.50	-0.15

Overall and Marsh (1980) pointed out, these differences should be viewed in the context of the 9-point Likert scale. Therefore, a difference of 0.19 or 0.15 is less than a quarter of a point on the Likert scale. Because the correlations were relatively strong (0.83 to 0.84) and the mean scores were similar, Overall and Marsh concluded that the results “provide strong support for the assumed stability of students’ evaluations of their courses and instructors” (p. 324). In other words, course evaluations, according to Overall and Marsh are a highly reliable measure.

Of particular importance in this study is the fact that the scores for each student were separated by a time period of one year or more. This is a long time lapse between administrations of the course evaluations and many things could have impacted the ratings students gave the course and instructor. It is surprising that the correlational coefficients were as high as they were (0.83 and 0.84).

Six years later, Scott and colleagues (1986) conducted a similar study. They collected data from medical students at the University of Washington, School of Medicine. Students in the study completed end of semester course evaluations and were then asked, both one year later and again two years later, to complete the same course evaluations for the same instructors. Thirteen courses were included in the study. The

evaluation used in the study asked students to answer two questions, using a 6-point Likert scale (1 = *very poor* to 6 = *excellent*): (a) Amount learned, and (b) Rating for the course. The ratings on item number 2 (the item most relevant to the present study) were statistically significantly lower, $F(12, 24) = 24.94, p < .000$, one year and two years later, than the original end of semester ratings for 12 of the 13 courses. Unfortunately, no mean ratings were reported in the article. Furthermore, the authors did not report the correlation between the two test administrations. The authors did, however, conclude that “this indicated that the original course ratings were not stable” (p. 765), or, in other words, that the ratings were not reliable.

When reliability in test-retest methodology is the subject under investigation correlational analyses should be used as the measure. Scott and colleagues (1986) failed to include correlational analyses in their study. Furthermore, they let one or two years pass between administrations of the course evaluation. Too much can occur in that amount of time that would affect the reliability of the measurement.

Later, West (1988) conducted a test-retest reliability study, and drew similar conclusions. Fifty-eight medical students from a regional university in the southeast United States participated in evaluating their instructors three times over the course of a single semester. The course evaluation employed consisted of eight items, seven of which were measured on a 5-point Likert scale; the eighth item asked students to provide an overall “grade” (A through F) of the instructor. A total of eight instructors were evaluated early in the course (time 1), again, six weeks later, in the second half of the course (time 2), and finally a third time, six weeks later, at the end of the course (time 3). West then conducted correlational analyses to examine the stability of the ratings over times 1, 2,

and 3. Table 3 presents both the median (from all eight items) correlation coefficients for each instructor, as well as the range of coefficients for the instructor. Additionally, an average-median correlation coefficient row has been added to help summarize the analyses that West conducted.

From the table it can be seen that median correlation coefficients for the eight instructors in the study ranged from a low of -0.05 to a high of 0.66. It can also be seen that the mean of the median coefficients calculated from the data collected between times 1 and 2 was 0.39, times 2 and 3 was 0.52, and between times 1 and 3 was 0.24. Overall these correlation coefficients were not very high. From these analyses, West (1988) concluded that the “results of this study do not support the argument that student evaluations of teachers remain consistent over the length of a course” (p. 111).

West (1988) used a much shorter, but still too long, time lapse between

Table 3

Median Correlation Coefficients for Instructor Ratings at Times 1, 2, and 3 from West (1988)

Instructor	Time comparison ^a					
	1 x 2		2 x 3		1 x 3	
	<i>r</i>	Range	<i>r</i>	Range	<i>r</i>	Range
1	.64	.46 - .75	.66	-.02 - .79	.53	.26 - .80
2	.35	-.09 - .56	.65	.28 - .83	-.05	-.20 - .31
3	.38	.20 - .47	.38	-.18 - .85	-.04	-.26 - .06
4	.35	.07 - .44	.66	.24 - .70	.42	-.14 - .55
5	.17	.04 - .49	.62	.52 - .79	.39	.03 - .63
6	.63	.01 - .77	.34	.13 - .48	.17	-.48 - .51
7	.35	.17 - .66	.47	.00 - .68	.26	.12 - .37
8	.23	.00 - .43	.35	-.14 - .58	.24	-.12 - .55
Median <i>r</i>	.35		.55		.25	

^a 1: early in the course; 2: in the second half of the course; 3: at the end of the course.

administrations of the course evaluations. Furthermore, he also used correlational analyses to look at the relationship between the ratings gathered at times 1, 2, and 3. These can be viewed as improvements over previous research.

Lastly, Witt and Burdalski (2003) conducted a test-retest reliability study of course evaluations using business students from a small Eastern United States university. A total of 14 courses were included in the study. Course evaluations at the university were collected 3 weeks before the end of the semester. The authors then collected ratings from the same students on the last day of the semester. The evaluation consisted of 10 items and were rated by students using a 5-point Likert scale (1 = *Well above average* to 5 = *Well below average*). Items 9 and 10 were of most importance for the present study: 9) Overall, how would you rate the effectiveness of this instructor? and 10) Overall, how would you rate your learning experience in this course? No correlational analyses were run on the collected data, but the authors did calculate “difference ratings” or the difference between the first and second administrations of the course evaluations. For items 9 and 10 the difference ratings were -0.05 and -0.03, respectively (neither was found to be statistically significantly different). According to the authors, this finding provided support for the conclusion that course evaluations do have short term stability or that they do provide a reliable measure.

Once again, the authors failed to include correlational analyses in their study. They did, however, improve on methodology employed in previous studies by administering the course evaluations at two points in time separated by only 3 weeks.

Table 4 is included to show a comparison of the four studies just described. A discussion of the different sections of the table follows.

Table 4

Comparison of Test-Retest Reliability Studies of Course Evaluations

Study	Sample size	Time lapse	Time of test administration	Focus of ratings	Items	Analyses		
						Correlations	Difference ratings	Authors conclusions
Overall & Marsh, 1980	1374	1 – 3 years	Last week of classes	Course	Single item	$r = 0.83$	0.15	Stable; reliable
				Instructor	Single item	$r = 0.84$	-0.19	
Scott, Hunt, & Greig, 1986	175	1 & 2 years	“End-of-course-ratings”	Course	Single item		$F(12, 24) = 24.94, p < .000$	Not stable; not reliable
West, 1988	234	6 weeks	Beginning, middle, and end of semester	Instructor	Single item	$r = 0.39$ $r = 0.52$ $r = 0.24$		Not stable; not reliable
Witt & Burdalski, 2003	306	3 weeks	3 weeks prior to end of semester; end of semester	Course Instructor	Single item Single item		-0.03 -0.05	Stable; reliable

Sample size. Researchers have included a varying number of students in their studies, ranging from 175 to 1,374 (three of the four studies included 306 or fewer). The authors of the study that included 1,374 students (Overall & Marsh, 1980), and the researchers who included 306 students (Witt & Burdalski, 2003), both concluded that course evaluations were stable and reliable; whereas the other two studies, which included 234 and 175 students both came to conclusions that course evaluations were not stable or reliable.

Time lapse. The studies included here employed a varying period of time between administrations of the course evaluations. Two of the studies administered the course evaluations with at least a year of time having elapsed (Overall & Marsh, 1980; Scott et al., 1986). This can be contrasted with West (1988) who waited 6 weeks between administrations and Witt and Burdalski (2003) who let only 3 weeks pass before

administering the course evaluations a second time. Again, the issue here is that test-retest reliability assumes that little has changed between the two administrations of the test. Particularly for course evaluations, this means that the students and their perceptions about the teacher have not changed or that they have not learned something new that affects the way they rate the instructor. It further applies to the instructor in that he or she has not changed his or her behavior in the classroom. For at least two (Overall & Marsh, 1980; Scott et al., 1986) of the four studies the assumption that nothing has changed is unlikely.

Focus of ratings. Both instructor and course ratings are of importance for this study as both give information about the effectiveness of instruction. Again, two studies (Overall & Marsh, 1980; Witt & Burdalski, 2003) included both course and instructor ratings, whereas of the other two studies, one (Scott et al., 1986) included only the course ratings and the other (West, 1988) included only ratings of the instructor.

Items. All of the measures used in the four studies were very similar to the course evaluation employed at USU, and based their analyses on single items. For example, in the Witt and Burdalski (2003) study both the course and instructor ratings were based on single items. For the instructor rating, students answered the following question: Overall, how would you rate the effectiveness of this instructor? To rate the course they were asked to answer the following: Overall, how would you rate your learning experience in this course?

It is interesting to note that only a very few studies have been done looking at the reliability of course evaluations, and none of those that have been conducted used more than a single item to determine the instructor or course rating. It is commonly known that,

all other things being equal, longer tests, that include more items, are more reliable than shorter tests. This is important because students are asked to take their time to complete the course evaluations. Evaluations with more items will consequently require more time of students. Therefore, a balance needs to be found that will provide good reliability, but not require too much time of students.

Analyses. To draw conclusions about the reliability of course evaluations it is important to have information about both the correlation and the differences in scores between the two administrations of the test. Two of the studies included in this analysis conducted correlational analyses (Overall & Marsh, 1980; West, 1988). Overall and Marsh found correlations of 0.83 and 0.84 on ratings of the course and the instructor, respectively. In the West study the average median correlations were 0.39 for time 1 x time 2, 0.52 for time 2 x time 3, and 0.24 for time 1 x time 3.

It is interesting to see that in the Overall and Marsh (1980) study, with one to three years between the administrations of the course evaluations, the coefficients were in the low 0.80 range. Whereas, the West (1988) study, with a time lapse of about 6 weeks (12 weeks between time 1 and time 3), average median correlation coefficients were a high of 0.52 and a low of 0.24. Once again, years or even 6 weeks time lapse between administrations of the course evaluation provide a lot of opportunity for changes to occur in the students, their learning and memory, and in the instructor.

The other two studies did not include correlational analyses. Scott and colleagues (1986) compared ratings from students at the end of the semester to ratings from 1 and 2 years later. They found that the ratings were statistically significantly lower 1 and 2 years later than they were on the course evaluations completed at the end of the semester.

Similarly, Witt and Burdalski (2003) compared means from course evaluations completed three weeks prior to the end of the semester with those from the end of the semester. They found that the means were not statistically significantly different. Lamentably, not enough data were provided by the authors to aid in calculations of correlation coefficients. Looking at the differences in means does not provide the proper reliability data like a correlational analysis does. The means from time 1 to time 2 could be very similar, but within the sample, individual data points at each time could be very different, some students rating an instructor high the first time and low the second time, with other students doing the reverse, rating the instructor low the first time and high the second time, while other students rate the instructor the same both times. This could produce two means that are very similar, but the correlation coefficient would be low.

Authors conclusions. Two of the four studies (Overall & Marsh, 1980; Witt & Burdalski, 2003) concluded that the course evaluations were stable and reliable. The authors from the other two studies (Scott et al., 1986; West, 1988) concluded that course evaluations are not stable or reliable.

Overall, Witt and Burdalski (2003) conducted their research using the most sound design: they included over 300 students, they focused on both the course ratings and the instructor ratings, and only had a three week time lapse between the first and second administrations of the course evaluations. The only problem in their study was that they did not include a correlational analysis of the relationship between the ratings from time 1 to time 2. Instead, they looked only at the differences in the means from time 1 to time 2. This is problematic because students could have changed their responses from time 1 to time 2, and the means could have still come out to be about the same.

A particular instrument used to measure teacher effectiveness would be deemed useless if it did not provide consistent, stable measures. Based on the four studies just reviewed, it is difficult to draw firm conclusions as to the reliability of course evaluations. For example, three of the four studies included time lapses between administrations of the course evaluations that were too long. Additionally, only two of the four studies included correlational analyses of the student ratings. The resulting coefficients for one study (Overall & Marsh, 1980) were high, but for the other study (West, 1988) they were very low. It is important, in test-retest methodology, to use a short time lapse and to include correlational analyses of the gathered data.

Reliability of measurement must first be established for conclusions regarding validity issues of course evaluations to be meaningful. If an instrument does not provide reliable measures of a particular construct, then it could not be concluded that it actually measures the construct it purports to measure. Validity is an important component of the psychometric properties of course evaluations. This issue will be discussed next.

Validity

In this context, validity, as it applies to course evaluations, refers to whether or not results of course evaluations can be used to make valid inferences about the construct they purport to measure (i.e., teacher effectiveness). Because there is no all encompassing, agreed upon definition of teacher effectiveness, many scholars have evaluated the validity of course evaluations by using student learning. They have posited that student learning is a measure of how effective the teacher has been in conveying the class material. Therefore, the validity of using student ratings as a measure of teacher

effectiveness has been estimated by examining the correlation between average student course ratings and average performance (student learning) in the course. In fact, two prominent scholars in the area of course evaluations have strongly emphasized the use of student learning as a measure of teacher effectiveness. In 1981 Cohen stated that “if student ratings are to have any utility in evaluating teaching, they must show at least a moderately strong relationship to student learning” (p. 281). Additionally, Marsh (2007) stated that “the most widely accepted criterion of effective teaching, appropriately, is student learning” (p. 338). This will be discussed next.

Student Learning

In courses where the teacher is effective, all else being equal, students should learn more than students in courses with teachers who are less effective. Based on this premise, researchers have evaluated the validity of students’ course evaluations by examining the correlation between student ratings and student performance in multisection courses in which each section is taught by a different instructor, but the same syllabus, textbook, and final exam (measure of student learning) is used. Of import in these studies, the level of analysis needs to be the average rating and average student learning (instead of individual students and ratings). Looking at individual students helps researchers answer the question of “whether students who learn more than other students, regardless of the class they are in, give higher ratings to instructors and courses” (Cohen, 1981, p. 284). Using average course ratings and average student learning allows researchers to differentiate the effects of instructor on student learning, and thereby address the issue of validity as it relates to course evaluations. As Kane, Gillmore, and

Crooks (1976) stated, “it is the dependability of the class means, rather than the individual student ratings, that is of interest, and the class is the appropriate unit of analysis” (p. 172).

In 2007, Marsh summarized the research on student ratings of teaching. In his report Marsh referenced two meta-analyses of the literature in the area of validity and student learning: Cohen (1981) and Feldman (1989). These two reviews were acquired. In addition, *Web of Science* was used to find articles that had cited the Feldman review (because it was the more recent of the two and because the Cohen [1981] review is too old to be included in *Web of Science*). The titles and abstracts of all 66 articles that cited Feldman (1989) were evaluated to determine if they were meta-analyses. None were found. Furthermore, it was found that the Feldman review did not present a meta-analysis of the multisection studies as the Cohen review did. Instead the Feldman review looked at comparing different sources of ratings, for example, ratings from students to self-ratings from the instructor. Therefore, *Google Scholar* and the “cited by” link were used to find all articles that had cited the Cohen meta-analysis. A total of 411 articles were found. The titles and abstracts from those articles were reviewed. Only one, a more recent meta-analysis, was located: *Student Evaluations of Teaching: Are They Related to What Students Learn?* (Clayson, 2009). Other articles that cited Cohen included discussions of faculty or administration opinions on course evaluations, ideas regarding using peer evaluation systems, or employing course evaluations for teaching improvement, to name a few. Both the Cohen and the Clayson meta-analyses will be discussed next.

In 1981 Cohen conducted a meta-analysis of 41 articles with a total of 68 multisection course studies (10 of the 41 articles included more than one multisection course).

To be included, these studies met the following criteria, as outlined by Cohen: (a) data presented had to be from actual college courses (no hypothetical data or data based on fabricated classroom situations), (b) the class had to be the unit of analysis, and (c) multisection courses, using a common final examination, had to be the basis of the study.

Cohen (1981) was interested in two relationships: (a) the relationship between overall rating of the instructor and student achievement and (b) the relationship between overall course rating and student achievement. The measure of student achievement used was a final examination score. The overall rating of both the instructor and the course was usually a single item for each. When more than a single item was used to measure overall rating, an average of all items included was calculated. Cohen did not report how many or which studies used more than a single item and had to be converted to an average. Next, the mean overall rating for both instructor and course was correlated with the mean student achievement score. These analyses will be discussed next.

Overall Course Rating with Student Achievement

Cohen (1981) states that only 22 of the total 68 courses provided data on overall course ratings. Therefore, only 22 courses were included in this analysis. In 20 (11 of which reported statistically significant correlations) of the 22 courses, the correlation between student achievement and overall course rating was positive. Furthermore, over half of the correlations fell at or above $r = 0.30$ (with a mean correlation of 0.47). Figure 1 presents the frequency of each correlation coefficient found in these analyses.

A total of 16 of the 22 courses (see Figure 1) had a correlation coefficient of 0.30 or higher. According to Cohen (1981), these data support the conclusion that there is a

positive relationship between overall course rating and student achievement. He explains that if there were no relationship then it would be expected that most coefficients would fall around zero with an equal number of positive and negative correlations. In this case, most (20 out of 22) were positive with a mean correlation coefficient of 0.47.

Overall Instructor Rating with Student Achievement

Sixty-seven (out of 68) courses provided the necessary data to calculate the correlation between overall instructor rating and student achievement. Fifty-nine of the 67 were positively correlated (29 of which were statistically significant). The overall mean correlation coefficient was $r = 0.43$. Figure 2 presents the frequency of the 67 correlations.

As can be seen in Figure 2, a total of 39 correlation coefficients fell at or above 0.5. Similar to the relationship between overall course rating and student achievement, according to Cohen (1981), the data from studies measuring the relationship between overall instructor rating and student achievement provide support for the conclusion that there is a positive relationship between student achievement and overall instructor rating: “instructors whose students achieved the most were also the ones who tended to receive the highest instructor ratings” (Cohen, 1981, p. 296).

These findings from Cohen (1981) regarding relationships between student achievement and overall instructor ratings supply some support for the notion that course evaluations can be used to make valid inferences about teacher effectiveness, when teacher effectiveness is operationally defined as “the degree to which an instructor facilitates student achievement” (p. 281).

Figure 1. Distribution of course/achievement correlations for 22 courses taken from Cohen, 1981.

Figure 2. Distribution of instructor/achievement correlations for 67 courses taken from Cohen, 1981.

More recently Clayson (2009) conducted a meta-analysis of multisection course validity studies of course evaluations. He was critical of past research in the area of course evaluations, stating that reviews in the area had only looked at studies published in “education related” journals or by those in “educational disciplines.” As an example he cited Cashin (1995) and the review he published. Clayson stated that Cashin’s review “consists of 67 references, only one of which came from a researcher outside of the educational disciplines” (p. 17). Additionally, but not mentioned by Clayson, Cashin only cited the Overall and Marsh (1980) study regarding the stability of course evaluations. When summarizing the research on validity, student ratings, and student learning, Cashin cited, just as Marsh (2007) did, Cohen (1981) and Feldman (1989).

Clayson (2009) stated that his meta-analysis is the first in the area of course evaluations to include studies from business journals. This is an important point for Clayson to make: in conducting a meta-analysis it is important for the researcher to find all studies that meet the inclusion criteria, and not to limit the search to only journals in a particular area.

For studies to be included in his meta-analysis, Clayson (2009) developed the following five criteria (very similar to those employed by Cohen [1981]): (a) “unit of analysis had to be directly related to college instruction,” (b) multiple sections of the same course had to be included, (c) “the measure of learning had to be common across all sections,” (d) the measure had to be a test, not something soliciting student perceptions, and (e) “the evaluation had to be conducted before the student completed the common

learning instrument” (p. 21). Based on these criteria, Clayson included 17 articles, which reported a total of 42 studies, in his meta-analysis.

Clayson (2009) reports that of the 42 studies included in his meta-analysis, 10 of them found a negative relationship between final exam scores and ratings on the course evaluations. Additionally, 13 found statistically significant positive relationships and 1 a statistically significant negative relationship. Correlation coefficients ranged from $r = -0.75$ to 0.91 (see Figure 3). This led him to conclude that “there does appear to be a positive association between learning and [ratings], but the average association’s magnitude is small [mean $r = 0.33$]” (p. 23).

This finding from Clayson (2009) is smaller than those findings from Cohen (1981), $r = 0.33$ as compared to $r = 0.47$ and $r = 0.43$. Overall, with 32 out of 42 correlation coefficients being positive and 13 statistically significant, these findings provide some support for the notion that course evaluations can be used to make valid inferences about teacher effectiveness.

A comparison was made (see Table 5) between those articles included in the Cohen (1981) meta-analysis and those included in the Clayson (2009) meta-analysis. Of Clayson’s 17, 10 were included in the Cohen (1981) meta-analysis. Of the seven not included in Cohen, five were published after 1981 and, therefore, were not available to Cohen. Appendix D contains a listing of all studies comparing those included in the Cohen meta-analysis to those included in the Clayson (2009) meta-analysis. Additionally, the table shows the results of a search for multi-section course validity studies not included in Cohen or Clayson; the search was conducted using *EBSCOHost*, *PsychInfo*,

Figure 3. Distribution of student rating/achievement correlations for 42 courses based on data presented by Clayson (2009).

and *Google Scholar* including the following keywords: *multi-section*, *course evaluation*, *student evaluation*, *student ratings*, and *validity*; one was found.

Of the seven studies included in Clayson (2009) that were not included in Cohen (1981) only two of them were published before 1981 and therefore could have been included in Cohen's meta-analysis. One study came from the *Journal of Educational Psychology*, whereas the other came from the *Journal of Economic Education*. This is interesting to note, because Clayson stated that he was the first to include studies from business journals in a meta-analysis of validity issues of course evaluations. He included one published article found in a business journal that Cohen failed to include in his meta-analysis. Of the other five included in the Clayson meta-analysis that were published after Cohen did his work, four came from business/economic journals. Therefore, it does appear that Clayson provided analyses of studies from outside the psychology/education

Table 5

Comparison of Studies Included in the Cohen (1981) and Clayson (2009) Meta-Analyses

		Clayson (2009) <i>n</i> = 17	
		Yes	No
Cohen (1981)	Yes	10	31
<i>n</i> = 41	No	7	1

Note. Appendix D presents a list of the studies included in both Cohen (1981) and Clayson (2009).

journals that had been previously overlooked. Interestingly, Clayson failed to include 31 studies that Cohen did include in his meta-analysis. Clayson does not offer any justification for this.

One study was found that was not included in either the Cohen (1981) or Clayson (2009) meta-analysis: *Student Evaluation of Lecturers as an Indicator of Instructional Quality: A Validity Study* (Orpen, 1980). This study will now be discussed.

Orpen (1980) reported that 128 students, from 10 sections (each taught by a different instructor) of a University Mathematics course, were included in the sample. The same textbook, content, assignments, and final examination were included in each section. Furthermore, there were 10 to 12 students enrolled in each section. The rating form employed consisted of 21 items, rated on a scale of 1 (*not at all descriptive of the lecturer*) to 5 (*very descriptive of the lecturer*). The ratings were summed to produce an overall rating of the instructor. Orpen then conducted a correlational analysis of the data, looking at mean final exam scores and the mean overall rating for the instructors. A result of $r = 0.52$, $p < .05$, prompted Orpen to conclude that, “it appears as if the student

evaluations were a highly accurate indication of the relative quality of instruction provided by the various lecturers” (p. 6). No means were reported by Orpen.

One problem in both meta-analyses, and the additional study, is the lack of reporting of mean ratings or mean final exam scores. This is a problem because mean ratings for instructors could be very highly correlated with mean exam scores from students in their courses, even though the differences in ratings or exam scores could be very large or very minimal. Without knowing the mean ratings and mean exam scores a meaningful conclusion about the findings from correlational analyses cannot be gained.

To demonstrate this, a Monte Carlo technique was conducted to create and analyze data from two hypothetical data sets, each consisting of five sections of a course, and each section containing 25 students. Ratings and exam scores were assigned to each student. Both data sets had a correlation coefficient of $r = 0.998$ between instructor ratings and exam scores. Data Set 1 had exam scores that ranged from 40 to 80 (mean = 60) and Data Set 2 had exam scores ranging from 79.96 to 80 (mean = 79.98). Specifics are presented in Table 6.

The two data sets are clearly very different and should lead to different conclusions. In Data Set 1, students in classes taught by instructors who were rated highly learned a lot more than students in classes taught by instructors who were rated low. In Data Set 2, students in classes taught by instructors who were rated highly learned virtually the same as the students in the classes taught by instructors who were rated low.

In both data sets, students rated the instructors exactly the same, but in Data Set 1 there is a large difference in the exam scores of students, whereas in Data Set 2 there is very little variation in students' exam scores. Following the logic of both Cohen (1981)

Table 6

Data from Hypothetical Data Sets 1 and 2 Looking at Validity of Course Evaluations

	Section	Mean instructor ratings	Mean exam scores	Correlation coefficient
Data Set 1				
	1	4.49	80.00	
	2	3.99	70.00	
	3	3.50	60.00	$r = 0.998$
	4	3.08	50.00	
	5	2.44	40.00	
Data Set 2				
	1	4.49	80.00	
	2	3.99	79.99	
	3	3.50	79.98	$r = 0.998$
	4	3.08	79.97	
	5	2.44	79.96	

and Clayson (2009) the correlation coefficients from either data set would lead to the conclusion that the student evaluations are a valid measure of instructor effectiveness because students in classes of highly rated instructors learn much more (i.e., students who learn more tend to be taught by more highly rated instructors). However, even though the resulting correlation coefficient in Data Set 2 is just as high as that from Data Set 1, the student exam scores are only trivially different. In this case, it does not make sense to say that student ratings of instructors are a strong predictor of student learning. Therefore, before concluding that positive correlations between teacher evaluations and student achievement provide evidence for the validity of student evaluations, it is important to know about the average exam scores in each class as well as the correlation coefficient. Unfortunately, neither Cohen nor Clayson provided any information about mean levels of achievement.

To investigate whether or not previous researchers had included these differences, an attempt was made to obtain all studies from both Cohen's (1981) and Clayson's (2009) meta-analyses. All but four of the studies included in Cohen's analysis were obtained (Crooks & Smock, 1974; Reynolds & Hansvick, 1978; Spencer & Dick, 1965; Wherry, 1952; all of which were unpublished studies). Of the obtained studies from Cohen and Clayson, it was found that three reported mean student ratings and final exam score means (Benton & Scott, 1976; Endo & Della-Piana, 1976; Frey, 1973). These means were included by the authors of these studies, but the authors did not analyze them or refer to them as being important. In other words, the means were reported in tables but no discussion of them being important for validity issues was included in the studies. Table 7 provides a comparison of the four studies (Frey, 1973, reported data from two studies). Additionally, Table 8 includes the means from each study. These studies are discussed below.

Focus of Rating

Course evaluations normally include two main foci, the instructor and the course. In the meta-analyses, studies that included data on both were included. Only the instructor was the focus of these four studies.

Subjects Included

Student participants have been included from a variety of subject areas in multisection designs. Math (specific course not specified), Introductory Calculus, Multidimensional Calculus, and Freshman English were courses included in these four

Table 7

Comparison of Studies Included in Cohen (1981) and Clayson (2009) That Included Means (ratings/final exams)

Study	Focus of rating	Subjects included	Number of sections	Number of students	Correlation coefficient
Benton & Scott, 1976	Instructor	Freshman English	31	265	0.22
Endo & Della-Piana, 1976	Instructor	Math	5	111	0.46
Frey, 1973	Instructor	Introductory Calculus	8	222	0.93*
	Instructor	Multidimensional Calculus	5	132	0.51

*Statistically significant at the .001 level.

studies. Frey (1973) included two studies, one with students from Introductory Calculus and one with students from Multidimensional Calculus.

Number of Sections

The number of sections included in these studies ranged from 5 to 31.

Number of Students

The number of students included in these studies ranged from 111 to 265. Only the Benton and Scott (1976) study included specific data on number of students per section (ranging from five to 13). In fact, Frey (1973) did not provide data regarding the number of students in each part of his study.

Table 8

Means Reported in those Studies Included in Cohen (1981) and Clayson (2009)

Benton & Scott, 1976		Endo & Della-Piana, 1976		Frey, 1973	
Final exam means	Instructor rating means	Final exam means	Instructor rating means	Final exam means	Instructor rating means
6.14	2.91	36.16	5.40	70.50	3.00
7.09	3.82	43.11	4.22	75.60	3.56
7.33	3.30	51.46	5.19	77.00	4.08
7.33	4.00	52.30	5.76	83.40	5.35
7.50	3.67	70.44	5.67	84.90	4.53
7.75	3.79		$r = 0.46$	85.60	5.04
8.00	4.04			87.10	5.10
8.11	3.26			89.20	5.16
8.33	2.50				$r = 0.93$
8.40	3.87				
8.43	3.57			62.90	3.67
8.62	3.72			71.20	3.81
8.72	4.15			71.50	5.55
8.78	4.11			74.00	4.37
9.00	2.37			77.80	4.74
9.00	2.72				$r = 0.51$
9.25	3.57				
9.40	4.03				
9.50	4.17				
9.73	3.59				
9.78	3.37				
9.78	3.70				
9.89	3.59				
10.00	3.73				
10.00	3.30				
10.33	3.81				
10.33	3.07				
11.50	3.54				
12.13	3.92				
12.14	4.24				
12.33	3.85				
	$r = 0.22$				

Correlation Coefficients

The resulting coefficients from the studies ranged from 0.22 to 0.93.

Analysis of the Means

An analysis of the relationship of the ratings and the final exam scores via correlations is necessary for understanding whether course evaluations provide a way of making valid inferences about instructor effectiveness. Also of importance in validity analyses is an analysis of the means. Table 8 presents the means reported by the three studies: Benton and Scott (1976), Endo and Della-Piana (1976), and Frey (1973).

As can be seen in Table 8, there is a reasonably good range of scores on the final exam scores. Without knowing the standard deviation of the distribution of individual student scores, it is impossible to know for sure, but it appears that in each study students in some classes learned substantially more than students in other classes. However, there are problems with how the final exam means and the mean ratings match up. For example, in the Benton and Scott (1976) study 9.00 is the median final exam mean. The lowest overall mean instructor rating was received by an instructor of a section of the course in which students scored at the median. Of the top 10 instructor ratings, five fell above the median final exam score and five fell below. Furthermore, the section with the highest overall mean final exam score (12.33), was rated number 10, with nine other instructors being rated higher. This is a scenario that would be expected with a correlation coefficient of $r = 0.22$.

In the Endo and Della-Piana (1976) study, five sections were included. The final exam means ranged from 36.16 to 70.44. The bottom two sections, based on final exam

scores, 36.16 and 43.11, rated their instructors 5.40 and 4.22, respectively. This means that the section with the lowest final exam mean, by almost 7 points (out of 100 total points), rated their instructor 1.22 points higher (on a scale of 1 to 7) than the students in the second lowest section. Additionally, the second highest section scored more than 18 points lower on the final exam than the top scoring section. But, the mean instructor rating for that section was higher (0.16 points) than the mean rating for the instructor in the highest scoring section. Again, this is consistent with a correlation coefficient of $r = 0.46$.

The correlation coefficient reported by Frey (1973), in his first study, is quite large ($r = 0.93$), which is consistent with the conclusion that the course evaluations do provide a way of making valid inferences about teacher effectiveness. By looking at the means from that Frey (see Table 8) study it can be seen that the final exam means and the mean instructor ratings line up almost perfectly, continually moving from lowest to highest.

Another important point in this analysis, as was demonstrated with the Monte Carlo study previously described in this literature review, is the differences between the final exam means. In Frey's (1973) second study, two sections scored almost identically, 71.20 and 71.50. If course evaluations do provide a way for making valid inferences regarding teacher effectiveness, then it would be expected that the mean instructor ratings for these two sections would be very similar as well. In reality, the ratings were very different. The students from the section that scored 0.30 points higher (out of 100 total points) on the final exam, actually rated their instructor 1.74 points higher (on a scale of 1 to 7). The mean instructor rating from this section also ended up to be the highest overall,

even though the final exam mean from the section was more than 6 points lower than the highest overall mean final exam score (77.80). Looking at means in this manner provides a better overall understanding of the resulting correlation coefficients and the relationship between final exam scores and ratings from course evaluations.

Utah State University Course Evaluation

The course evaluation (see Appendix E) employed at Utah State University, has been in use for at least the last 10 years. It contains four sections, each composed of a differing number of items. Section 1 contains two items: (a) The overall quality of the course was... and (b) The instructor's effectiveness in teaching the subject matter was... Section 2 is composed of eight items focused on "Information about the course." Section 3 contains 10 items about "Information about the instruction." Lastly, Section 4 is composed of five items about "Information about students."

Very little previous research has been conducted on the USU course evaluation. Two available studies were found. The first was a dissertation (Olson, 2008) looking at the two open-ended questions on the backside of the evaluation form. The second, pertinent to the present discussion, available report of research was the 2010 Faculty Evaluations Committee Annual Report to the Faculty Senate Executive Committee (USU, 2010). In the report a Cronbach's alpha analysis, conducted by Dr. Jamison Fargo, of all available completed course evaluations ($n = 50,962$) from the Fall 2008 semester was presented. Table 9 presents the Cronbach's alpha results.

As a measure of internal consistency, Cronbach's alpha is based on correlations between the items on the course evaluation. It provides an understanding of the degree to

which the items in an instrument tend to measure the same construct. As can be seen in Table 9, the items that make up each section of the USU course evaluation have high internal consistency. Internal consistency can be contrasted with test-retest reliability in that test-retest reliability seeks to measure the consistency of the measurement from one time to another.

Cashin (1995) reported that the consistency or inter-rater agreement (Cronbach's alpha) on course evaluations differs depending on the number of students who complete the evaluation. He states that for an instrument with 38 items, the result, based on 10 students would be $r = 0.69$; $r = 0.83$ for 20 students; $r = 0.88$ for 30 students; and, $r = 0.91$ for 40 students. He goes on to suggest that results from measures with reliability coefficients below $r = 0.70$ should "be interpreted with particular caution" (p. 2).

Conclusions

Based on the review of the literature, it can be seen that more information is needed before drawing conclusions like those provided by recognized experts in the area of course evaluations (see Table 1). For example, Cashin (1995) stated that course

Table 9

*Results from Cronbach's Alpha Analysis
of the Items from the Utah State University
Course Evaluation*

Items	Cronbach's Alpha
Section 1: 1 & 2	0.92
Section 2: 1 – 8	0.95
Section 3: 1 – 10	0.96

evaluations “tend to be statistically reliable” (p. 6). Surprisingly, only four test-retest reliability studies could be found and those reliability studies reached contradictory conclusions. Of those four studies only one was cited by Cashin (1995). Furthermore, results from the four reliability studies failed to include a reasonable time lapse between administrations of the course evaluations, and two of the four only examined differences in mean scores over time instead of correlations of results after a short time span.

Moreover, other experts have asserted that course evaluations are the “most valid source of data on teaching effectiveness” (McKeachie, 1997, p. 1219; Scriven, 1988). Based on the present literature review, more information is needed before making conclusions similar to those offered by experts. Based on the information currently available, a multisection course validity study needs to be conducted in which the means (of ratings and exam scores) are also analyzed and compared. This will offer a better understanding of whether or not course evaluations (similar to the one employed at USU) provide a reasonable means for making valid inferences regarding teacher effectiveness.

Overall, the results of the current study will provide evidence for or against the conclusions previously drawn by experts in the area of course evaluations (see Table 1).

Therefore, the present study consisted of two parts: a test-retest reliability study and a multisection validity study.

CHAPTER III

METHOD

Based on the review of the literature, the study was designed to answer the following research questions:

1. What is the test-retest reliability of the USU course evaluation from time 1 (3 weeks before the end of the semester) to time 2 (the end of the semester)?
2. What is the test-retest reliability of the USU course evaluation from time 2 (the end of the semester) to time 3 (3 weeks after the semester has ended)?
3. What is the test-retest reliability of the USU course evaluation from time 1 (three weeks before the end of the semester) to time 3 (3 weeks after the semester has ended)?
4. Can the results of the course evaluation employed at USU be used to make valid inferences about a teacher's effectiveness?

After data were collected for the study, it was determined that the response rate for time 3 was very low. Because this group was likely not representative of the students who typically complete course evaluations, results for research questions 2 and 3 will be reported but not discussed. Therefore, the two research questions for the present study are:

1. What is the test-retest reliability of the course evaluation currently employed at Utah State University (USU)? and
2. Can the results of the course evaluation employed at USU be used to make valid inferences about a teacher's effectiveness?

Test-Retest Reliability Design

The test-retest design part of the current study will help provide a better understanding of the reliability of the course evaluation used at USU, thereby answering the first research question.

Participants

Students from the following courses, during the Spring 2010 semester, were included in the test-retest study: Introductory Psychology (PSY 1010; three sections) and Intermediate Algebra (MATH 1010; five sections). Instructors of the courses were contacted and asked if their students could be invited to participate. The classes were then visited by the researcher and the study was presented to them, asking for their participation.

Materials

The current course evaluations (see Appendix E) at USU as well as the letter of information (see Appendix F) were employed in this study. Also, a self-addressed stamped envelope, including a blank course evaluation, was provided to students during the second administration of the course evaluations.

Procedure

Each class was visited three weeks before the end of the semester (time 1) and the letter of information (see Appendix F) was read to students. Students were then asked to complete the standard course evaluation form currently used at USU (see Appendix E). Each student was asked to write the last 4 digits of their student identification number on

the evaluation. Then, 3 weeks later, at the end of the semester (time 2) the official course evaluations were administered as they are normally, with the following difference: the researcher accompanied the departmental secretary to the classes, reminded students of the study, and informed them that they again needed to write the last four digits of their identification number on the evaluation. The secretary, once the course evaluations had been administered, took copies of the completed course evaluations and provided the copies to the researcher.

During the second administration of the course evaluations, the researcher provided students with the self-addressed stamped envelope. Adhered to the outside of the envelope was a note reminding students of the date they needed to complete the enclosed course evaluation and mail it back to the researcher. Additionally, email addresses were collected from students. The email addresses were used to remind students to complete the course evaluation the third time and mail it back to the researcher. Two days prior to the date students were to return the course evaluation a reminder email was sent to each student. Then, one week later, a second reminder email was sent.

Multisection Course Design

Conducting a multisection course validity study will help provide a better understanding of the degree to which the results of course evaluations at USU can be used to make valid inferences about a teacher's effectiveness, thus answering the second research question.

Intermediate Algebra (MATH 1010) at USU has many sections each semester. For example, there were 16 sections of MATH 1010 during the Spring 2010 semester. Those 16 sections used the same textbook and the same final exam, but there were 11 different instructors for the 16 sections. Final exam score data from MATH 1010 from Fall 2005 through Spring 2010 were used to do a multisection course validity study of student evaluations. Mean ratings from the course evaluations completed at the end of each semester, from Fall 2005 through Spring 2010, along with mean scores on the final exam from those semesters, were collected and analyzed (correlations and differences between means).

Participants

Existing data, with no identifying information attached to them, were used. Permission was obtained from the head of the Mathematics department to access the existing mean final exam scores for the past 5 years of the MATH 1010 course. Data were unavailable for a few sections.

Student ratings data are made public each semester through the office of Analysis, Assessment, and Accreditation (AAA) on USU campus (see the Facts and Figures Website, here: <http://aaa.usu.edu/p&a/CourseEval/FacEval.htm>). The data corresponding to the MATH 1010 sections were collected through this website.

Materials

Course evaluations. The current course evaluation at USU, which has been in use for at least the past 10 years (see Appendix E) was employed in this part of the study. The first two items on the course evaluations: (1) *The overall quality of this course was...*

and (2) *The instructor's effectiveness in teaching the subject matter was...*, were the intended focus of the analyses (following Cohen's, 1981 analysis). Additionally, subsequent analyses also examined data in which average scores for multiple items dealing with course quality and instructor effectiveness were calculated. Students used a 6-point rating scale in answering all items, ranging from 1 = *very poor* to 6 = *excellent*, on the course evaluation.

Final exam. The final exam employed by the Mathematics department at USU for MATH 1010 provided the mean student score data needed for this part of the study. Unfortunately, no reliability data were available for the final exams. At the end of each school year, based on student performance, the final exam is changed by a faculty committee in the Math department. Therefore, the data included in the present study were analyzed by year.

Procedure

Existing data, including mean final exam scores from MATH 1010 for the past five years, were collected from the Math department. Table 10 presents the semester, total number of sections taught, and the number of sections of available data.

There were three reasons why data were not available for all sections. First, the Mathematics department could not find the needed data (this was the case for 28 sections). Second, the data were aggregated by semester or by instructor (this was the case for the 17 sections in Fall 2006), Third, some sections during a single semester were taught by the same instructor (i.e., one instructor might teach three sections). In these

Table 10

Number of Sections of MATH 1010 and the Available Data from the Last Five Years

Semester	MATH 1010 Sections taught	Data available (MATH 1010 Sections)
Spring 2010	16	11
Fall 2009	20	16
Spring 2009	17	11
Fall 2008	20	12
Spring 2008	13	10
Fall 2007	18	11
Spring 2007	13	8
Fall 2006	17	0 (only pooled data available)
Spring 2006	13	10
Fall 2005	18	11

instances, one of the sections for the instructor was randomly selected to be included in the study.

Lastly, the mean student ratings for the corresponding MATH 1010 sections were collected, via the Analysis, Assessment, and Accreditation office website at USU.

CHAPTER IV

RESULTS AND DISCUSSION

The results from each part of the study will be discussed separately, starting with the test-retest reliability part, followed by the multi-section course validity part.

Test-Retest Reliability Design

Once the data had been collected, descriptive statistics were calculated. As shown in Table 11, this included sample sizes, means, and standard deviations (see Table 12).

At each data collection time students were asked to put the last four digits of their identification number on the course evaluation. Overall, about 11.6% of the students did not do this; therefore the total number of unique individual students who participated in this part of the study cannot be determined. For example, a student may have recorded her identification number during the first administration of the course evaluations, but may have failed to do so during the second administration. In this case there would be no way of connecting her first course evaluation to her second course evaluation, and she would show up in the data as two different individuals. When this happened, the data from that individual were not included in the correlational analyses, but were still used in the overall descriptive statistics data.

Based on the data from Table 11, there were a total of 514 students in data collection Time 1 (46 who did not include the last four digits of their identification number), a total of 579 students in data collection Time 2, (188 who did not include the last four digits of their identification number), and 50 total students in data collection Time 3 (all of whom included the last four digits of their identification number). These

Table 11

Sample Sizes from the Course Sections Included in the Test-Retest Reliability Study

Instructor	Enrollment	Sample size (no ID number in parentheses)		
		Time 1	Time 2	Time 3
A	51	22 (5)	30 (2)	8 (0)
B	55	25 (2)	29 (14)	4 (0)
C	498	278 (21)	322 (125)	19 (0)
D	51	28 (2)	27 (8)	5 (0)
E	248	116 (10)	112 (31)	6 (0)
F	102	45 (6)	59 (8)	8 (0)
	1005	514 (46)	579 (188)	50 (0)

Table 12

Mean Ratings and Standard Deviations from the Course Sections Included in the Test-Retest Reliability Study

Instructor	Mean instructor rating (standard deviation in parentheses)			Mean course rating (standard deviation in parentheses)		
	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
A	4.00 (1.16)	4.00 (1.11)	4.13 (1.25)	4.95 (2.01)	4.33 (1.24)	4.38 (1.19)
B	4.92 (0.86)	4.83 (0.97)	5.50 (0.58)	5.56 (0.96)	5.28 (0.75)	5.75 (0.50)
C	5.50 (0.77)	5.66 (0.58)	5.74 (0.45)	5.64 (0.78)	5.76 (0.56)	5.89 (0.31)
D	5.21 (0.69)	5.41 (0.69)	5.60 (0.55)	5.54 (1.26)	5.59 (0.57)	5.40 (0.89)
E	4.54 (0.97)	4.87 (0.98)	5.00 (0.89)	4.44 (1.10)	4.84 (1.14)	4.50 (0.84)
F	5.58 (0.66)	5.66 (0.58)	5.88 (0.35)	5.67 (0.56)	5.83 (0.38)	6.00 (0.00)

data have also been separated by instructor. Mean ratings of the course per instructor ranged from 4.44 to 6.00. Mean ratings of instructor ranged from 4.00 to 5.88.

It is interesting to note the number of students, per instructor, who completed course evaluations at each time (see Table 11). Often, fewer than half the number of enrolled students in each participating course completed the course evaluation in this study. For example, Instructor A had a total enrollment of 51 students. At Time 1 only 22 students completed the course evaluation. Instructor E had a total enrollment of 248 students. A total of 116 course evaluations were completed at Time 1 and only 112 at Time 2. This raises concern about the representativeness of the sample included in the present study. This will now be discussed.

There was about a 12% increase in the number of students who completed the course evaluations from Time 1 to Time 2 (Time 2 was the official USU course evaluation). Even with this increase, the problem remains that only about half of the students enrolled completed the official course evaluation. For instance, in the course taught by Instructor B, with an enrollment of 55, the number of students who completed the course evaluation went from 25 at Time 1 to 29 at Time 2. Similarly, for Instructor F, with a total enrollment of 102, there were 45 completed evaluations at Time 1 and an increase to Time 2 to 59 completed course evaluations.

To better understand the lack of completion, by students, of the course evaluation at USU, 18 on-campus courses, from the Spring 2010 semester, were randomly selected from three different colleges: the College of Agriculture, the College of Humanities, Arts, and Social Sciences, and the College of Science. Four of the courses were selected from general education (labeled as 1010 or 1020 in the course catalog: Biology 1010, English 1010, Nutrition and Food Science 1020, and Sociology 1010) courses, seven from upper division (labeled between 3000 and 5000 in the course catalog: Biology 3100, English

3310 and 4420, Nutrition and Food Science 4560 and 4720, and Sociology 3120 and 3500) courses, and seven from graduate level (labeled above 5000 in the course catalog: Biology 5150 and 5250, English 6350, Nutrition and Food Science 6410 and 7800, and Sociology 6630 and 7100) courses. Total enrollment for the course during the Spring 2010 semester was collected (based on data available on ACCESS). In addition, the number of students who completed the official course evaluation at the end of the semester (data available on the AAA Office website) was also collected. Lastly, the average ratings (for both instructor and course) were collected for each course. Table 13 presents the collected data.

These data show that the higher the enrollment in the course, the lower the percentage of students who complete the course evaluation ($r = -0.40$). For example, in general education courses with an enrollment of over 100 students, just over 56% of the students complete the course evaluation. In contrast, the general education courses with an enrollment below 100 have an average of 70.5% of students who complete the course evaluation. Furthermore, the upper division and graduate level courses (with an average enrollment of $n = 20$) have an average of 87.4% of the student enrollment who complete the course evaluation.

Based on these findings it can be concluded that the completion rate in the present study was not outside the norm at USU for the larger, over 100 enrollment general education course (PSY 1010). The completion rate for the smaller general education course, MATH 1010, does appear to be lower than the USU norm. Therefore, even though general education courses were included in an effort to get a more representative sample of undergraduate students at USU, this may not have been the case.

Table 13

Course Evaluation Completion Data from Spring 2010 Semester at Utah State University

Course	Section number	Total enrollment	Course evaluations completed	Instructor rating	Course rating
General education courses					
Biology 1010	1	277	161 (58.1%)	4.3	4.3
	2	192	103 (53.7%)	5.1	5.0
English 1010	1	20	10 (50%)	5.5	5.2
	2	22	16 (72.7%)	4.9	5.0
	3	19	5 (26.3%)	5.0	5.2
	4	19	12 (63.2%)	5.7	5.6
	6	22	15 (68.2%)	5.1	5.1
	7	22	18 (81.2%)	5.6	5.4
	8	18	14 (77.8%)	5.1	5.2
	9	22	19 (86.4%)	4.5	4.6
	10	22	14 (63.7%)	5.4	5.6
	11	21	18 (85.7%)	5.7	5.6
	12	22	16 (72.7%)	5.6	5.5
	13	22	20 (90.9%)	5.3	5.1
	14	21	10 (47.6%)	5.6	5.4
	15	22	15 (68.2%)	5.5	5.5
	16	22	15 (68.2%)	4.7	4.9
	17	23	18 (78.3%)	5.5	5.3
	18	24	14 (58.3%)	5.3	5.3
	19	22	19 (86.4%)	5.5	5.5
	20	22	20 (90.9%)	5.5	5.5
Nutrition & Food Science (NFS) 1020	1	468	276 (59.0%)	4.1	4.2
Sociology 1010	1	243	122 (50.2%)	5.2	5.1
	2	244	149 (61.1%)	4.0	4.3
	3	75	55 (73.3%)	5.0	4.9
Upper division courses					
Biology 3100	1	22	19 (86.4%)	5.3	5.2
English 3310	1	28	25 (89.3%)	5.5	5.3
English 4420	1	19	16 (84.2%)	5.3	5.4
NFS 4560	1	17	17 (100%)	3.1	3.5
NFS 4720	1	16	16 (100%)	5.8	5.4
Sociology 3120	1	31	26 (83.9%)	5.0	5.2
Sociology 3500	1	52	39 (75.0%)	5.3	5.3
Graduate level courses					
Biology 5150	1	25	17 (68.0%)	2.8	3.5
Biology 5250	1	26	19 (73.1%)	5.7	5.6
English 6350	1	12	11 (91.7%)	5.6	5.6
NFS 6410	1	2	2 (100%)	5.5	5.5
NFS 7800	1	11	9 (81.8%)	4.6	4.8
Sociology 6630	1	6	6 (100%)	5.8	5.7
Sociology 7100	1	10	9 (90.0%)	6.0	6.0

Next, the Pearson Product Moment correlation coefficients were calculated to determine the relationships of the following: (a) overall quality of the course ratings received at data collection Time 1 and data collection Time 2, (b) instructor's effectiveness ratings received at data collection Time 1 and data collection Time 2, (c) overall quality of the course ratings received at data collection Time 2 and data collection Time 3, (d) instructor's effectiveness ratings received at data collection Time 2 and data collection Time 3, (e) overall quality of course ratings received at data collection Time 1 and Time 3, and (f) instructor's effectiveness ratings received at data collection Time 1 and Time 3. The resulting coefficients are presented in Table 14.

As can be seen in Table 14, the correlation coefficients range from a high of 0.94, for the relationship between course ratings from Times 1 and 3, to a low of 0.64, for the relationship between course ratings from Times 1 and 2. All six were statistically significant at the .01 level. There was a very low response rate for data collection Time 3. Only 50 students returned completed course evaluations. As shown in Table 14, the

Table 14

Test-Retest Reliability Correlation Coefficients

	Sample size (<i>n</i>)	Correlation coefficient
Instructor ratings		
Time 1 x Time 2	265	0.74
Time 2 x Time 3	44	0.85
Time 1 x Time 3	39	0.86
Course Ratings		
Time 1 x Time 2	261	0.64
Time 2 x Time 3	44	0.93
Time 1 x Time 3	39	0.94

Note. All coefficients were statistically significant at the .01 level.

overall course ratings received at Time 1 and Time 3 were most highly correlated ($r = 0.94$), with the correlation coefficient for the same ratings from Time 2 and Time 3 being very similar ($r = 0.93$). In other words, for the small subset of students who returned evaluations forms at Time 3, the results of student course evaluation are quite consistent with their evaluations at Times 1 and 2. However, given that correlations between Time 1 or 2 and Time 3 could only be computed for 39 to 44 students, it is very unlikely that these students are representative of the 265 students for whom some data are available. Maybe the students who returned the completed course evaluations during the third data collection time are more dedicated or perhaps they are more studious or perhaps they felt more committed to the instructor than the students who did not return the course evaluation. The fact is that it is unknown how these students differed from the larger group, but because they are such a small group and because the results for this group are so radically different from the other results, the data from Time 3 are not included in subsequent analyses and discussion.

Determining the importance of these test-retest correlations (all were statistically significant at the .01 level) is not easy to do. Charter (2003) stated that “it is difficult to establish standards for retest reliability because many factors need to be considered...” (p. 292). The correlation coefficients in the present study were statistically significant, but more specific standards for reliability coefficients applicable to the present context need to be discussed.

Charter (2003) conducted an analysis of reliability data gathered from “2,733 test critiques, eight journal articles, and 47 test manuals” (p. 293). He found a total of 933 reliability coefficients, 439 of which were test-retest correlation coefficients. Table 15,

Table 15

Descriptive Statistics for Retest Reliability Coefficients for Ten Test Types Taken from Charter (2003)

Test	Mean	SD	Percentile			n
			25	50	75	
Clinical adult	.74	.18	.64	.79	.86	36
Clinical child	.77	.13	.73	.80	.86	41
Aptitude	.81	.08	.76	.81	.89	12
Personality	.79	.10	.71	.81	.86	64
Vocational	.76	.15	.68	.81	.85	46
Attitude	.78	.14	.67	.78	.89	19
Neuropsychological	.85	.11	.76	.87	.94	27
Intelligence	.80	.14	.71	.86	.90	36
Educational	.79	.13	.72	.81	.87	35
Achievement	.78	.20	.68	.86	.95	11
Other	.80	.12	.73	.81	.91	112
Total	.79	.13	.71	.81	.86	439

taken from Charter (2003), presents the findings from those 439 coefficients, broken down into 10 different test types (and an *Other* category). Mean coefficients, standard deviations, value of coefficient at the 25th, 50th, and 75th percentiles, and the total number of coefficients per each test type are also included.

Data from this table provide a better understanding of the context within which reliability data from the course evaluations employed at USU may be placed. The reliability coefficient of $r = 0.64$ (course rating for Time 1 and 2) obtained in the current study falls at or below the 25th percentile for every type of test included in Charter's (2003) table. The coefficient from the instructor ratings at Time 1 and Time 2 in the current study, $r = 0.74$, falls below the median for every type of test when compared to the data from Charter (2003). The course evaluation at USU does not provide consistent, stable measures over a three-week time period.

Cronbach's alpha analyses were also conducted on the data from the present study from both Time 1 and Time 2. This provides a way of comparing this course evaluation data to that employed in the analyses conducted by Dr. Fargo (USU, 2010). Table 16 presents the results as compared to those from Dr. Fargo.

The results from the present study are similar to those from Dr. Fargo and would therefore, warrant a conclusion that the data from the present study are representative of student responses on the course evaluation at USU.

Because course evaluations are also used in promotion and tenure decisions, further explication of the context for understanding these results will now be presented. A hypothetical example could be the following: assume that if an instructor's rating is in the bottom 10% on the course evaluation, he will be terminated. If the course evaluation were a perfectly reliable measure and had a mean of 100 and a standard deviation of 20, then an instructor with a score below 74 (-1.28, standard score for bottom 10%, multiplied by the standard deviation of 20 below the mean) would be terminated. Because every measure has some error, for example, with a 0.90 reliability coefficient or a 0.70 reliability coefficient, then making decisions based on those

Table 16

*Comparison of Cronbach's Alpha Using Data from
the Present Study and from USU (2010)*

Items	Time 1	Time 2	Dr. Fargo
Section 1: 1 & 2	0.87	0.89	0.92
Section 2: 1 – 8	0.90	0.93	0.95
Section 3: 1 - 10	0.95	0.95	0.96

measures will be inaccurate to some degree. For example, due to the error in the measurement some instructors may receive ratings above the cutoff point of 74 that would have been below the cutoff if no error were included in the measure (false negative, should have been terminated but was not). Other instructors may receive ratings above the cutoff in both cases (true positive, should not have been terminated and was not). Still other instructors may receive ratings below the cutoff, but without error in the measurement would have received ratings above the cutoff (false positive, should have been retained but was terminated). Finally, some instructors would receive ratings below the cutoff in both cases (true negative, should have been terminated and was).

Using such data (mean, standard deviation, reliability coefficients), Charter and Feldt (2001) explained the results of such a hypothetical example with 0.90 and 0.70 reliability coefficients in the context of providing intervention to clients. Their results have been placed in the present context of retaining or terminating instructors, but the numbers have been unchanged. With a reliability coefficient of 0.90 only 7.8% of instructors would correctly be terminated, 2.2% would be terminated but should have kept their jobs, and 2.2% would keep their jobs who should have been terminated. Lastly, 87.8% would keep their jobs who should have kept them (see Table 17).

Table 17

*Retaining and Terminating Instructors Based
on a Reliability Coefficient of 0.90*

		Observed scores	
		Retained	Terminated
True scores	Retained	87.8%	2.2%
	Terminated	2.2%	7.8%

Charter and Feldt (2001) go on to explain that with a 0.70 coefficient the numbers change as follows: 6.0% correctly terminated, 4.0% retained and should have been terminated, 4.0% terminated who should have been retained, and 86.0% retained who should have been retained (see Table 18). Charter and Feldt conclude their explanations by stating, “when test scores have a low reliability the observed (fallible) scores are poor indices of true scores. However, when the reliability is high the observed scores can accurately predict the true scores” (p. 534).

This data will now be discussed in the context of USU. At USU, as of the Fall 2009 semester, there were a total of 684 tenure-track faculty positions; keeping with the bottom 10% being terminated, that would mean that about 68 faculty members would be let go. If the course evaluations were 100% reliable and the decision to terminate or retain a faculty member was based solely on the ratings received on course evaluations, then the decision could be made with confidence. Based on the data from the present study, it can be seen that this is not the case. A reliability coefficient of 0.64 for course rating and 0.74 for instructor rating was found. Therefore, employing the highest reliability coefficient ($r = 0.74$ from instructor rating), a hypothetical data set was constructed and analyzed using methods similar to those presented by Charter and Feldt (2001). The resulting

Table 18

*Retaining and Terminating Instructors based
on a Reliability Coefficient of 0.70*

		Observed scores	
		Retained	Terminated
True scores	Retained	86.0%	4.0%
	Terminated	4.0%	6.0%

percentages are presented in Table 19.

It can be seen in Table 19 that the resulting percentages are very close to those presented by Charter and Feldt (2001) for a reliability coefficient of $r = 0.70$. Therefore, keeping in mind the 684 tenure-track faculty positions at USU, there would be a total of 25 (3.7% of 684) faculty members who would be terminated who should have been retained. Further, there would be 25 retained who should have been terminated. Additionally, based on observed scores, about 37% of those who would be terminated should have been retained, and about 37% of those who were retained should have been terminated.

Based on the resulting data and the preceding discussion, further analyses were conducted. The second section of the USU course evaluation contains eight questions about the course (see Appendix E). It can be seen that these questions, taken together, could be used to give an overall rating for the course. Therefore, a mean rating was calculated for data that included responses on all eight questions. Then a correlation between the average score at Time 1 and the average score at Time 2 was computed, with

Table 19

Hypothetical Example of Using Course

Evaluations with a Reliability Coefficient of

0.74 to Terminate Instructors at Utah State

University

		Observed scores	
		Retained	Terminated
True scores	Retained	86.3%	3.7%
	Terminated	3.7%	6.3%

a resulting correlation coefficient of $r = 0.73$, $p = .00$.

In addition, the third section of the course evaluation contains 10 questions that ask about the instruction provided by the teacher. Taken together, these questions could be used to give an overall rating for the instructor. Therefore, a mean rating was calculated for student responses that included responses on all 10 questions. Then a correlation between the average score at Time 1 and the average score at Time 2 was computed, with a resulting correlation coefficient of $r = 0.78$, $p = .00$.

Following these two analyses, the same procedure used previously, looking at faculty who would be retained and those who would be terminated, was conducted. The results are displayed in Table 20.

In comparing the data in Table 19 to the data in Table 20, it can be seen that the percentages changed very little. The change would have an effect of subtracting two faculty members from the groups of those being terminated who should have been retained and those being retained who should have been terminate. Additionally, about a third of those who were terminated should have been retained and about a third of those who were retained should have been terminated.

Table 20

Hypothetical Example of Using Course Evaluations

with a Reliability Coefficient of 0.78 to Terminate

Instructors at Utah State University

		Observed scores	
		Retained	Terminated
True scores	Retained	86.7%	3.3%
	Terminated	3.3%	6.7%

Lastly, a comparison was made between the first correlational analyses conducted on the overall course rating (single item) and instructor rating (single item) questions and the second analysis that included means from all questions in the second and third sections of the course evaluation (multiple items). Table 21 presents the results.

As can be seen in Table 21, there was an increase in the resulting test-retests reliabilities for both the course rating and the instructor rating when aggregated data were used. The reliability coefficient for course rating went from 0.64 to 0.73, and the reliability coefficient for instructor rating went from 0.74 to 0.78. These coefficients are an improvement, but still not high enough to use for making important decisions about an individual's performance.

Of particular importance is that these results are the same whether or not the course evaluations are actually measuring teaching effectiveness. This point of validity will be addressed in the next section of this chapter.

Table 21

Comparison of Reliability Coefficients for Overall- and Mean-Course and Instructor Ratings

Analysis	Sample size	Reliability coefficient
Overall course rating (single item)	261	0.64
Mean course rating (multiple items)	243	0.73
Overall instructor rating (single item)	265	0.74
Mean instructor rating (multiple items)	255	0.78

The major focus of the course evaluations, both by promotion and tenure committees and by the AAA Office, is the first section or the questions that ask students to rate the overall course and instructor. In fact, the AAA Office does not even publish the data from the other sections of the course evaluations. Taken alone, these individual questions are not a very reliable measure. Using a mean rating from the second and third sections of the course evaluation provides a slightly higher reliability coefficient, 0.73 and 0.78 for course and instructor, respectively. Reliability coefficients at this level are better, but they are still too low.

Multisection Course Design

The final exam used for the MATH 1010 sections at USU is changed each school year based on the results from the previous year. Therefore, all collected data were separated by both semester and school year. They were then analyzed as follows: first, correlation coefficients were calculated for: (a) final exam means and mean instructor ratings, and (b) final exam means and mean course ratings (see Table 22); second, the differences in the means were analyzed.

Table 22 displays the overall final exam means as well as the associated mean instructor and course ratings for each semester included in the previous analyses. (The single-item ratings for both course and instructor had to be used in this part of the study because they were the only data available from 2005 to 2010; data from sections 2 and 3 of the course evaluations were not available.) For means separated by individual instructors see Appendix G. The results of the correlational analyses show a wide range of correlation coefficients, by both semester and year, between the mean final exam

scores and both the instructor and course ratings, coefficients range from -0.39 to 0.71.

Of the 18 individual semester correlation coefficients, six were negative. The only statistically significant correlation coefficient resulted from the data in the Spring 2007 semester, between mean final exam score and overall course rating ($r = 0.71$). The resulting correlation coefficients by year (excluding those from just the Spring 2007 as the data for the Fall 2006 semester were unavailable) ranged from -0.06 to 0.27, with only one negative and no statistically significant correlations.

Table 22

Multisection Course Validity Study Descriptive Statistics and Correlation Coefficients

					Correlation coefficients			
					By semester		By year	
Semester	Number of sections (<i>n</i>)	Final exam means (<i>SD</i>)	Mean instructor ratings (<i>SD</i>)	Mean course ratings (<i>SD</i>)	Exam x instructor rating (<i>r</i>)	Exam x course rating (<i>r</i>)	Exam x instructor rating (<i>r</i>)	Exam x course rating (<i>r</i>)
Fall 2005	11	120.45 (6.60)	5.10 (0.76)	4.97 (0.61)	0.09	0.07	0.04	-0.06
Spr 2006	10	126.74 (10.11)	5.14 (0.54)	4.97 (0.46)	-0.02	-0.20		
Spr 2007	8	129.79 (9.67)	5.16 (0.70)	5.11 (0.56)	0.64	0.71 ^a	NA	NA
Fall 2007	11	126.64 (7.80)	4.26 (1.01)	4.30 (0.87)	0.03	0.14	0.09	0.11
Spr 2008	10	132.31 (6.89)	5.01 (0.78)	4.89 (0.56)	-0.23	-0.39		
Fall 2008	12	121.08 (9.94)	5.22 (0.47)	5.13 (0.47)	-0.10	-0.07	0.02	0.06
Spr 2009	11	121.27 (7.82)	5.09 (0.52)	4.98 (0.44)	0.19	0.25		
Fall 2009	16	108.51 (7.70)	4.75 (0.83)	4.71 (0.73)	0.41	0.32	0.27	0.15
Spr 2010	11	134.29 (5.23)	4.96 (0.84)	4.77 (0.83)	0.23	0.14		
				Mean correlation	0.14	0.11	0.11	0.07

Note. Statistically significant at the .05 level; NA = Correlation coefficients for year-long data are not available.

There could be a few reasons for the resulting low correlational coefficients in this part of the study: (a) The variables may not be linearly related, (b) the correlations could have been attenuated by an extreme outlier, and (c) the MATH 1010 final exam may have had low reliability. To address the first two issues, scatterplots for data from each semester, and for both course rating and instructor rating, were created (see Appendix H). The scatterplots do not show any particular non-linear relationships between the variables or outliers that could have dramatically lowered the correlation coefficient.

Interestingly, in the scatterplots (for both Instructor and Course ratings) for the Spring 2007 semester, which had the highest correlation coefficient of $r = 0.71$ for Course rating, it appears that a single data point was responsible for the high correlation. That data point (exam mean: 116.0, Course Rating: 4.0) was removed and the analysis was run again. The resulting correlation coefficient fell from $r = 0.71$ to $r = 0.50$. This result shows how much influence a single outlier data point can have on correlation coefficients, especially when there are a relatively small number of data points.

Based on the data available the reliability of the MATH 1010 final exam can not be addressed. No data exist on the psychometric properties of the final examination employed by the Mathematics department for their MATH 1010 course.

The data presented in Table 22 show some similarities and some differences to the data presented in the meta-analyses of multisection course validity studies conducted by Cohen (1981) and Clayson (2009). Of the studies included in the Cohen analysis, 24 included less than 10 sections and 19 included between 10 and 19 sections. Of those included in Clayson, 16 included less than 10 sections and 12 included between 10 and

19 sections. In the present study, each semester included data for 10 sections or more, except Spring 2007, which included data for 8 sections.

Additionally, both meta-analyses found both positive and negative correlation coefficients with a large range. In contrast to both Cohen (1981), who reported average correlations of $r = 0.47$ and $r = 0.43$, and Clayson (2009), who reported an average correlation of $r = 0.33$, the average correlational coefficients in the present study were $r = 0.14$ and $r = 0.11$ for exam by instructor rating and exam by course rating respectively. The data from the current study support conclusions from Clayson that, “there does appear to be a positive association between learning and [student ratings of instructors], but the average association’s magnitude is small and inconsistent” (p. 23). There is no clear pattern to the resulting coefficients, with a third of them falling below zero and just over 72% of them falling below $r = 0.30$.

Because it is possible to have high correlations when mean scores are only trivially different it is important to look at the mean scores in the data used in multisection course validity studies. Looking at the data from Appendix G it is clear that there is quite a bit of range in final exam scores. Consistent with the low correlations, there are many examples of high exam scores being paired with low student evaluations and vice versa. For example, in the Spring 2006 semester data, the instructor who had the lowest final exam mean (114.36) had the same mean course rating (5.3) as the instructor who had the highest final exam mean (143.21). Additionally, in the Spring 2008 semester data it can be seen that the instructor who had the second highest final exam mean (138.21) had an instructor rating that was 1.7 points lower than the instructor who had the lowest final exam mean (120.79).

CHAPTER V

CONCLUSIONS

The present study was conducted in an effort to answer the following research questions:

1. What is the test-retest reliability over a three-week period of the course evaluation currently employed at Utah State University? and,
2. Can results of the course evaluation employed at Utah State University be used to make valid inferences about a teacher's effectiveness?

Expert Opinion

In the effort to answer these questions, it was found that expert opinion regarding the psychometric properties of course evaluations (see Table 1) was consistent, but not supported by research. For example, Cashin (1995) stated that course evaluations “tend to be statistically reliable, valid, and relatively free from bias or the need for control; probably more so than any other data used for evaluation” (p. 6). Additionally, McKeachie (1997) concluded that course evaluations “are the single most valid source of data on teaching effectiveness” (p. 1219).

Based on the review of literature, it can be seen that more information and data are needed before drawing conclusions similar to these offered by experts. In the area of reliability, a proper time lapse needed to be employed as well as correlational analyses of the resulting data. Three of the four test-retest reliability studies that had been conducted failed to include a reasonable time lapse between administrations of the course

evaluations, and two of the four failed to use correlational analyses. In the area of validity a better understanding, such as an analysis of the mean scores, is needed before making conclusions, like those from the experts, that they are “not only a valid, but often the only valid, way to get much of the information needed for most evaluations—and for many useful descriptions—of teaching” (Scriven, 1988, pp. 14-15).

It can be inferred, from previous research, that the course evaluations used in that research were similar to the USU course evaluation. For example, all of the test-retest reliability studies found conducted analyses utilizing single-rating items. Additionally, Cohen (1981) reported that most of the studies included in his meta-analysis used a single item to rate the instructor and/or the course (no specific data were provided as to the number that included only a single item). Clayson (2009) did not mention this aspect of the studies included in his meta-analysis. This is important because it provides an understanding of how the results from the present study can be generalized to other institutions and other course evaluations.

Response Rate

The fact that course evaluations at USU are not completed by many students raises serious questions about the representativeness of the data. For example, it is fairly common across all six instructors who were part of the present study, that less than half of the enrolled students completed the course evaluations. Furthermore, there does appear to be, in the larger general education courses at USU, a completion rate of around 60%, in upper division undergraduate course a completion rate of about 70%, and in graduate level courses a completion rate of just over 80%. Therefore, the completion rate in the

present study does not appear to be too far away from the norm at USU. These results do provide insight into attendance rates or student attitudes toward course evaluations in some courses at USU. They also raise issues of use of the course evaluations. If so few students are actually completing the course evaluation, do the ratings received by instructors reflect their actual teaching effectiveness? Why did so few students complete the course evaluation? Why do fewer students complete the course evaluation in courses with higher enrollment? What motivates students to complete the course evaluation? What do students think about and consider while they complete the course evaluation?

Test-Retest Reliability

In answering the first research question posed in this study, what is the test-retest reliability over a three-week period of the course evaluation currently employed at Utah State University?, it can be seen that the reliability coefficients in the current study were lower than is necessary to conclude that the course evaluation employed at USU provides a reliable measure. After using a three-week time lapse between administrations of the course evaluations, all correlation coefficients (Time 1 x Time 2 for both instructor and course ratings) were statistically significant at the .01 level, and ranged from $r = 0.64$ to $r = 0.74$.

It is encouraging that there was an increase in the coefficients when a mean overall course rating and instructor rating were calculated from the second and third sections of the USU course evaluations. Unfortunately, this increase was not sufficient to draw conclusions that the course evaluations employed at USU provide a stable measure

over time. The results provide support for the idea that the current course evaluation at USU needs to be revised to provide a more reliable measure.

Validity

This part of the study sought to answer the second research question, can results of the course evaluation employed at USU be used to make valid inferences about a teacher's effectiveness? Resulting correlation coefficients, between final exam means and both course and instructor ratings, fell between -0.06 and 0.27, with a mean of $r = 0.11$ for instructor rating and $r = 0.07$ for course rating. Such low correlations can be explained in part by the mediocre reliability scores reported in the previous section. However, such low correlations also suggest that even if reliability could be improved, the current course evaluation questions do not differentiate between effective and ineffective instructors.

Previous research had failed to include or analyze the differences in the means from the exams and the ratings. Only three previous studies had included these means, but had done so without considering or analyzing them. The present study was conducted to replicate previous research in an effort to consider and analyze the means. This provides a better understanding of the significance of the differences in ratings from one instructor to another, but this would only be the case when the resulting correlation coefficients support the conclusion that course evaluations do provide a means of making valid inferences about the teaching effectiveness of those being evaluated.

The present findings from the multisection course validity part of the study provide more evidence for the notion that course evaluations do not provide a means for

making valid inferences about teaching effectiveness. Further, they tend to support the second half of Clayson's (2009) conclusion that there appear to be inconsistencies in the findings from multisection course validity studies. In comparison to reports of previous multisection course validity studies, average correlation coefficients of $r = 0.14$ and $r = 0.11$, for instructor and course ratings, respectively, were found in the present study. Cohen (1981) found average correlation coefficients of $r = 0.44$ for instructor rating and $r = 0.47$ for course rating; whereas Clayson found an average correlation coefficient of $r = 0.33$.

The data and results from the present analyses do not offer support for the conclusions drawn by experts in the area of course evaluations (see Table 1). For instance, based on the results from the present study, the conclusion must be made that the course evaluation currently employed at USU are not reliable. Basing decisions, like promotion and tenure, retaining or terminating instructors, on the ratings from the course evaluations would lead to many wrong decisions. In addition, the number of students who complete the course evaluations is a cause of concern. Additionally, the results of the present study provide no evidence that the course evaluation employed at USU provides means for making valid inferences about the teaching effectiveness of instructors. There is no evidence that the most effective instructors receive the highest ratings from students, or that the least effective instructors receive the lowest ratings.

Limitations

The first limitation to the current study was that it was conducted at USU and only included the course evaluation form currently employed at the university. This was

purposeful. As was discussed previously, the course evaluation currently in use at USU is similar to the course evaluations that have been included in previous test-retest reliability and multisection course validity studies. Therefore, findings from the present study are relevant to those institutions that use similar course evaluations. Future research could include different universities or could include different types of course evaluations.

A second limitation in the present study was that only students from introductory courses at USU were included in the present study. This was done in an effort to collect data from a representative sample of undergraduate students. Both MATH 1010 and PSY 1010 are general education requirements and include a broad range of the student population at USU. The response rate received during the study matched the response rate of students in large undergraduate courses at USU. Therefore, the findings from this study may not be applicable to upper division courses or even to graduate level courses. Future research could include students from those populations. However, because undergraduate students constitute the largest part of the university and because most instructors teach some undergraduate courses, the results of this study are important.

Test-Retest Reliability

Limitations in this part of the study include the loss of students from data point 1 (three weeks before the end of the semester) to data point 2 (end of the semester), as well as the lack of student response during data point 3 (three weeks after the end of the semester).

Many students failed to include the last four digits of their identification number during administrations of the course evaluations. In order to connect students during each

administration, they were asked to place the last four digits of their identification number on the course evaluations. A total of 231 students across all three administrations did not include this piece of data. Having access to other variables, such as gender, SAT score, and attendance rate, would have provided a better way of comparing students who included their ID number with those who did not include their ID number. Therefore, this issue cannot be completely addressed, but a t test was run comparing the completed course evaluations from both groups. This was done with those from data collection time 1 and time 2. The groups were compared on both instructor rating and course rating. Table 23 provides the results of the analyses.

As can be seen in Table 23, no statistically significant differences were found between the groups, at either time or on either rating. Therefore, there is no evidence that students who were included in the present study were dramatically different in how they rated instructors than those who were excluded (based on lack of the last four digits of their identification number).

Table 23

Differences in Data from Students Who Included Their ID Number and Those Who Did Not

Data collection	Rating	Number of students		Means		<i>t</i> statistic	Statistically significant?
		ID included	No ID	ID included	No ID		
Time 1	Course	469	44	5.20	4.98	-1.23	No
	Instructor	463	42	5.30	5.00	-1.58	No
Time 2	Course	396	181	5.35	5.40	0.73	No
	Instructor	395	180	5.48	5.45	0.75	No

Of all students who completed at least one course evaluation during the study, only 50 returned a completed course evaluation three weeks after the semester ended (time 3). Additionally, only 37 total students completed all three course evaluations. This lack of response is troubling. Future research that asks students to return evaluations after the semester has ended needs to find ways of convincing more students to participate.

Multisection Course Validity

In this part of the study one possible problem could have been the unavailability of all the data for all sections of the MATH 1010 course over the past five years. An entire semester of data was unavailable (17 sections). In addition, several sections each semester were not available (a total of 30 sections over the 5-year period). Sometimes this was due to the data not being archived. Other times it was because the available data were aggregated (i.e., all data from one instructor who taught three sections), and therefore could not be included in the analyses. This could potentially change the results of correlational analyses, increasing or decreasing the resulting coefficient.

Implications

The results from the current study suggest that the current course evaluation employed at USU is not a reliable measure. Including multiple items to assess course effectiveness or instructor effectiveness, instead of focusing on a single item for each does improve the reliability coefficients (from $r = 0.74$ to $r = 0.78$ and from 0.64 to 0.73 for instructor rating and course rating, respectively). However, these increases are so small that it really does not solve the problem.

Also of importance is the use of course evaluations in promotion and tenure decisions. There are obviously other things that are incorporated into the decision made by a promotion and tenure committee, but the data from the study suggest that including data from the presently employed course evaluation at USU could frequently lead to incorrect decisions due to the unreliability of the results.

The results from the present study also raise questions about the use of the USU course evaluation for making valid inferences about teacher effectiveness.

Obviously, promotion and tenure decisions about individual faculty members, helping instructors improve their teaching, and providing students with information about instructors when they register for courses are important parts of higher education. The student ratings that result from course evaluations have been used as a major element in these activities. Results from the present study raise serious concerns regarding the continued use of course evaluations similar to the one in use at USU.

Lastly, based on the low response rate, low reliability coefficient, and the low association between student learning and the student ratings from the USU course evaluation, future research needs to focus on both student and faculty/administration attitudes and perspectives on course evaluations. Perhaps students do not consider course evaluations to be important or that no one does anything with the results. On the other hand, maybe faculty and administrators do not help students realize the importance of the evaluations.

REFERENCES

- Bejar, I. I. (1975). A survey of selected administrative practices supporting student evaluation of instruction programs. *Research in Higher Education*, 3(1), 77-86.
- Benassi, V. A., & Seidel, L. F. (2006). Using student evaluations to improve teaching. In W. Buskist & S.F. Davis (Eds.), *Handbook of the teaching of psychology* (pp. 279-284). Malden, MA: Wiley-Blackwell.
- Benton, S. E., & Scott, O. (1976). *A comparison of the criterion validity of two types of student response inventories for appraising instruction*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA. Retrieved from <http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED128397>
- Cashin, W. E. (1988). *Student ratings of teaching: A summary of the research*. IDEA Paper No. 20. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED302567&site=ehost-live>
- Cashin, W. E. (1989). *Defining and evaluating college teaching*. IDEA Paper No. 21. Retrieved from http://www.theideacenter.org/sites/default/files/Idea_Paper21.pdf
- Cashin, W. E. (1995). *Student ratings of teaching: The research revisited*. IDEA Paper No. 32. Retrieved from http://www.theideacenter.org/sites/default/files/Idea_Paper_32.pdf
- Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco, CA: Jossey-Bass.

- Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of General Psychology, 130*(3), 290-304.
- Charter, R. A., & Feldt, L. S. (2001). Meaning of reliability in terms of correct and incorrect clinical decisions: The art of decision making is still alive. *Journal of Clinical and Experimental Neuropsychology, 23*(4), 530-537.
- Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn?: A meta-analysis and review of the literature. *Journal of Marketing Education, 31*(1), 16-30.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.
- Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*(3), 281-309.
- Crooks, T. J., & Smock, H. R. (1974). *Student ratings of instructors related to student achievement*. Unpublished paper, Urbana, Illinois: Office of Instructional Resources, University of Illinois.
- Endo, G. T., & Della-Piana, G. (1976). A validation study of course evaluation ratings. *Improving College and University Teaching, 24*(2), 84-86.
- Feldman, K. (1989). The association between student-ratings of specific instructional dimensions and student-achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30*(6), 583-645.

- Frey, P. W. (1973). Student ratings of teaching: Validity of several rating factors. *Science*, 182(4107), 83-85.
- Kane, M. T., Gillmore, G. M., & Crooks, T. J. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement*, 13(3), 171-183.
- Kolitch, E., & Dean, A. V. (1999). Student ratings of instruction in the USA: Hidden assumptions and missing conceptions about 'good' teaching. *Studies in Higher Education*, 24(1), 27-42.
- Lin, Y., McKeachie, W. J., & Tucker, D. G. (1984). The use of student ratings in promotion decisions. *The Journal of Higher Education*, 55(5), 583-589.
- Marsh, H. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R.P. Perry & J.C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319-383). The Netherlands: Springer.
- Marsh, H. W. (1977). The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. *American Educational Research Journal*, 14, 587-595.
- Marsh, H. W., & Overall, J. U. (1981). The relative influence of course level, course type, and instructor on students' evaluations of college teaching. *American Educational Research Journal*, 18(1), 103-112.
- McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist*, 52(11), 1218-1225.

- McKeachie, W., & Svinicki, M. (2005). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (12th ed.). Belmont, CA: Wadsworth.
- Olson, D. (2008). Teaching patterns: A pattern language for improving the quality of instruction in higher education settings. *All Graduate Theses and Dissertations*. Retrieved from <http://digitalcommons.usu.edu/etd/51>
- Orpen, C. (1980). Student evaluation of lecturers as an indicator of instructional quality: A validity study. *The Journal of Educational Research*, 74(1), 5-7.
- Overall, J. U., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology*, 72(3), 321-25.
- Penny, A. R. (2003). Changing the agenda for research into students' views about university teaching: Four shortcomings of SRT research. *Teaching in Higher Education*, 8(3), 399-411.
- Pietrzak, D. (2010). *IDEA student survey of teaching and courses: Examination of score consistency*. Report to the Office of Evaluation and Assessment, University of South Dakota.
- Reynolds, D. V., & Hansvick, C. (1978, September). *Graduate instructors who grade higher receive lower evaluations by students*. Paper presented at the annual meeting of the American Psychological Association, Toronto, Ontario.
- Scott, C. S., Hunt, D. D., & Greig, L. M. (1986). Changes in course ratings following clinical experiences in the clerkship years. *Academic Medicine*, 61(9), 764-766.
- Scriven, M. (1988). The validity of student ratings. *Instructional Evaluation*, 92(2), 5-18.

- Seldin, P. (1975). How colleges evaluate professors. Current policies and practices in evaluating classroom teaching performance in liberal arts colleges. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED111282&site=ehost-live>
- Seldin, P. (1984). Faculty evaluation: Surveying policy and practices. *Change*, 16(3), 28-33.
- Seldin, P. (1989). How colleges evaluate professors. 1988 vs. 1983. *AAHE Bulletin*, 41(7), 3-7.
- Seldin, P. (1993). The use and abuse of student ratings of professors. *The Chronicle of Higher Education*, 39(46), A40.
- Spencer, R. E., & Dick, W. (1965). *Course evaluation questionnaire: Manual of interpretation*. Champaign, IL: Office of Instructional Resources, University of Illinois.
- Utah State University. (2010). *Faculty evaluations committee annual report to the faculty senate executive committee*. Logan, UT: Office of the Provost.
- West, R. F. (1988). The short-term stability of student ratings of instruction in medical school. *Medical Education*, 22(2), 104-112.
- Wherry, R. J. (1952). *Control of bias in ratings* (PRS Reports 914, 915, 919, 920, and 921). Washington, DC: Department of the Army, The Adjutant General's Office.
- Witt, J., & Burdalski, D. (2003, September). *Regarding the timing of student course/instructor evaluations*. San Francisco, CA: Academy of Business Educators Conference.

APPENDICES

Appendix A:
List of Institutional Research Centers

Institutional Research Centers from 50 States

^Δ Arizona State University	*University of Chicago
^Δ Arkansas State University-Jonesboro	^Δ University of Delaware
Boise State University	University of Hawaii
^Δ Brown University	^Δ University of Maryland
Cleveland State University	*University of Michigan
^Δ Colorado State University	University of Mississippi
Florida State University	University of Missouri
^Δ Georgia University of Technology	^Δ University of Montana
*Iowa State University	University of Nebraska-Lincoln
Jacksonville State University	*University of Nevada-Las Vegas
Kansas State University	^Δ University of New England
Kentucky State University	*University of New Hampshire
Louisiana Tech University	^Δ University of New Haven
Minnesota State University-Mankato New	University of North Dakota
Mexico State University	University of Oklahoma
*New York University	^Δ University of South Carolina
^Δ North Carolina State at Raleigh	*University of South Dakota
^Δ Oregon State University	*University of Utah
*Pennsylvania State University	University of Vermont
Rutgers	*University of Wisconsin-Madison
*Purdue University	University of Wyoming
*Texas A&M University	Virginia Polytechnic Institute and State
*Tennessee State University	University
Tufts University	Washington State University
*University of Alaska-Anchorage	West Virginia University
University of California-Davis	

All Institutions listed were sent the first email soliciting reliability and validity research on course evaluations that had been conducted by their center.

*Replied via email

^ΔContacted via phone

Appendix B:
Email to Offices of Institutional Research

Director,

My name is Mike Drysdale and I am a graduate student in Psychology working on a dissertation. I am looking at the reliability and validity of course evaluations. As I have examined the literature in this area I have only been able to find 4 test-retest reliability studies and very few multi-section validity studies. I am assuming that this is only the case in the published literature and that there is other research out there that just has not been published. In an effort to be more comprehensive I am inquiring with centers like yours to see if they have conducted reliability and validity studies of course evaluations. If your center has conducted such studies could you email me a copy of the report or let me know how I can obtain it?

Thank you very much for your time.

Mike

Appendix C:
Follow-up Email to Offices of Institutional Research

Dear Director/President,

I am just following up with my previous request, regarding research conducted at your institution on the reliability and validity of course evaluations. I have received several responses so far. About 80% of those responses were that no known research in this area had been done at their institution. The other 20% have responded with articles/reports that have been helpful in my search. I've added below a list of some of the relevant research that I have found and included in my analyses. These are what I've been able to find thus far. If you are aware of other research in this area please let me know, particularly if it has been done at your institution and perhaps was not published.

Clayson, D. E. (2009). Student Evaluations of Teaching: Are They Related to What Students Learn?: A Meta-Analysis and Review of the Literature. *Journal of Marketing Education*, 31(1), 16-30.

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

Marsh, H. (2007). Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness. In *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective* (pp. 319-383).

Overall, J. U. & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology*, 72(3), 321-25.

Scott, C. S., Hunt, D. D., & Greig, L. M. (1986). Changes in course ratings following clinical experiences in the clerkship years. *Academic Medicine*, 61(9).764-766.

West, R. F. (1988). The short-term stability of student ratings of instruction in medical school. *Medical Education*, 22(2), 104-112.

Witt, J. & Burdalski, D. (2003, September). Regarding the timing of student course/instructor evaluations. Academy of Business Educators Conference, San Fransisco, CA.

Thank you for your help. I very much appreciate it.
Mike

Appendix D:

Table of Studies Included (and those not included) in Cohen (1981) and Clayson (2009)

Article	Cohen (1981)	Clayson (2009)
Bendig (1953a)	X	X
Bendig (1953b)	X	
Benton & Scott (1976)	X	
Bolton, Bonge, & Marr (1979)	X	
Borg & Hamilton (1956)	X	
Braskamp, Caulley, & Costin (1979)	X	X
Bryson (1974)	X	
Centra (1977)	X	X
Chase & Keene (1979)	X	
Cohen & Barger (1970)	X	
Costin (1978)	X	X
Crooks & Smock (1974)	X	
Doyle & Crichton (1978)	X	
Doyle & Whitely (1974)	X	X
Elliott (1950)	X	
Ellis & Rickard (1977)	X	
Endo & Della-Piana (1976)	X	
Frey (1973)	X	X
Frey (1976)	X	
Frey, Leondard, & Beatty (1975)	X	X
Greenwood, et al. (1976)	X	
Grush & Costin (1975)	X	
Hoffman (1978)	X	
Johnson (2003)		X
Marsh, Fleiner, & Thomas (1975)	X	
Marsh & Overall (1980)	X	
McKeachie, Lin, & Mann (1971)	X	
Mintzes (1977)	X	
Morsh, Burgess, & Smith (1956)	X	
Murdock (1969)	X	
Orpen (1980)		
Palmer, Carliner, & Romer (1978)		X
Rankin (1965)	X	
Remmers, Martin, & Elliott (1949)	X	
Reynolds & Hansvick (1978)	X	
Rodin & Rodin (1972)	X	X
Rubinstein & Mitchell (1970)	X	
Sheets, Topping, & Hoftyzer (1995)		X
Shmanske (1988)		X
Solomon, Rosenberg, & Bezdek (1964)	X	
Soper (1973)		X
Sorge & Kline (1973)	X	
Spencer & Dick (1965)	X	
Sullivan & Skanes (1974)	X	X
Turner & Thompson (1974)	X	
Weinber, Fleisher, & Hashimoto (2007)		X
Wherry (1952)	X	
Whitely & Doyle (1979)	X	
Yunker & Yunker (2003)		X

Appendix E:
USU Course Evaluation

UtahState
UNIVERSITY

TEACHER/COURSE EVALUATION

INSTRUCTOR: _____ **COURSE:** _____ **SECTION:** _____

Student evaluations are an important part of the assessment of teaching effectiveness. Please respond as honestly and candidly as possible. Disregard questions which do not seem to be applicable. The completed forms and the computer data will not be available to the instructor until after class grades are awarded.

Please use black or blue pen or no. 2 pencil only.

EXCELLENT VERY GOOD GOOD FAIR POOR VERY POOR NOT APPLICABLE

I. GENERAL EVALUATION

1. The overall quality of this course was:	☐	☐	☐	☐	☐	☐
2. The instructor's effectiveness in teaching the subject matter was:	☐	☐	☐	☐	☐	☐

II. INFORMATION ABOUT THE COURSE

1. The extent to which course objectives were clear was:	☐	☐	☐	☐	☐	☐	☐
2. Relevance of assignments to course content was:	☐	☐	☐	☐	☐	☐	☐
3. Relevance of material presented in class to course goal(s) was:	☐	☐	☐	☐	☐	☐	☐
4. Appropriateness of workload to course goal(s) was:	☐	☐	☐	☐	☐	☐	☐
5. Relevance of exams to course goal(s) was:	☐	☐	☐	☐	☐	☐	☐
6. Fairness of course grading procedures was:	☐	☐	☐	☐	☐	☐	☐
7. The extent to which course responsibilities of students were clarified was:	☐	☐	☐	☐	☐	☐	☐
8. Helpfulness of assigned text/readings to achieving course goal(s) was:	☐	☐	☐	☐	☐	☐	☐

III. INFORMATION ABOUT THE INSTRUCTION

1. The extent to which course organization helped learning was:	☐	☐	☐	☐	☐	☐	☐
2. The helpfulness of explanations by the instructor, if/when needed, was:	☐	☐	☐	☐	☐	☐	☐
3. Instructor's use of examples, if/when appropriate, was:	☐	☐	☐	☐	☐	☐	☐
4. Instructor's use of class time to help students learn the subject matter was:	☐	☐	☐	☐	☐	☐	☐
5. Instructor's enthusiasm for subject of course was:	☐	☐	☐	☐	☐	☐	☐
6. Instructor's helpfulness in resolving student's questions was:	☐	☐	☐	☐	☐	☐	☐
7. The extent to which the instructor was prepared for class was:	☐	☐	☐	☐	☐	☐	☐
8. Opportunity to ask questions was:	☐	☐	☐	☐	☐	☐	☐
9. Opportunity for students to make comments and express opinions was:	☐	☐	☐	☐	☐	☐	☐
10. Availability of extra help, if/when needed, was:	☐	☐	☐	☐	☐	☐	☐

IV. INFORMATION ABOUT STUDENTS

- At the beginning of the semester, my interest in the subject matter of the course was: ☐ High ☐ Medium ☐ Low
- My current GPA at USU is in the range of: ☐ 4.0-3.5 ☐ 3.4-3.0 ☐ 2.9-2.5 ☐ 2.4-2.0 ☐ 1.9-1.0
- This course is being used for: ☐ my major ☐ my minor ☐ a liberal Arts & Sciences major, minor or certificate
☐ general education ☐ an elective ☐ other
- I am a: ☐ freshman ☐ sophomore ☐ junior ☐ senior ☐ graduate ☐ other
- Grade I expect to receive is: ☐ A ☐ B ☐ C ☐ D ☐ F ☐ Pass

Please answer the two questions on the back.

Appendix F:
Letter of Information

Letter of Information
Title: Psychometric Properties of Course Evaluations

Introduction/Purpose: Dr. Karl White in the department of Psychology at Utah State University (USU) and Mike Drysdale are conducting a research study to find out more about the psychometric properties of course evaluations. There will be approximately 300 participants in this research.

Procedures: If you agree to be in this research study you will be asked to complete the USU course evaluation once, 3 weeks before the semester ends; the official course evaluation that you fill out at the end of the semester will also be included in the study; and to complete the course evaluation a third time 2 weeks after the semester ends.

Risks: There are minimal risks involved in the study.

Benefits: There are unlikely to be any direct benefits to you from these procedures. The investigator, however, may learn more about course evaluations.

Explanation & offer to answer questions: Mike Drysdale has explained this research study to you and answered your questions. If you have other questions or research-related problems, you may reach him at (435) 797- 7278.

Compensation: There is no foreseeable immediate compensation for you participating in this research study, only a knowledge that you are helping further our understanding of the evaluation of teachers.

Voluntary nature of participation and right to withdraw without consequences: Participation in research is entirely voluntary. You may refuse to participate or withdraw at any time without consequences or loss of benefits. Your decision not to participate will not affect your class standing.

Confidentiality: Research records will be kept confidential, consistent with federal and state regulations. Only the investigator and assistants will have access to the data which will be kept in a locked file cabinet in a locked room. Information will be kept for 3 years, at which point it will be destroyed. Only the last 4 digits of your A# will be requested; please do not put your name on the course evaluations.

IRB Approval Statement: The Institutional Review Board for the protection of human participants at USU has approved this research study. If you have any questions or concerns about your rights, you may contact the IRB at (435) 797-1821.

Investigator Statement: “I certify that the research study has been explained to the individual, by me, and that the individual understands the nature and purpose, the possible risks and benefits associated with taking part in the research study. Any questions that have been raised have been answered.”

Karl R. White, Principal Investigator

(435) 797-3013

Michael Drysdale, Co-Principal
Investigator

(435) 797-7278

Appendix G:
Final Exam Means and Instructor/Course Ratings Separated by Instructor

Appendix H:
Scatterplots for Final Exam Mean x Course Rating and Final Exam Mean x Instructor
Rating

Final Exam Mean x Course Rating
Fall 2005 ($r = 0.07$)

Spring 2006 ($r = -0.20$)

Final Exam Mean x Instructor Rating
Fall 2005 ($r = 0.09$)

Spring 2006 ($r = -0.02$)

CURRICULUM VITAE

Michael J. Drysdale

Psychology Department
2810 Old Main Hill
Logan, UT 84322-2810

Office: ECERC 310A
Phone: 435-764-0548
Email: mikedrys@gmail.com

Education

PhD in Psychology (Research & Evaluation Methodology), Utah State University,
December 2010

MS in Psychology (Research & Evaluation Methodology), Utah State University,
September 2009

BA, major in Psychology, minor in Spanish, Southern Utah University, May 2004

Teaching Experience

- Psychometrics (PSY 5330; taught via interactive broadcast)
 - Fall 2008
- Scientific Thinking & Methods in Psychology (PSY 3500; taught via interactive broadcast)
 - Fall 2006
- Psychological Statistics (PSY 2800; taught via interactive broadcast)
 - Spring 2008
- Introductory Psychology (PSY 1010):
 - Fall 2009
 - Spring 2008
 - Fall 2007 (2 sections)
 - Spring 2007
 - Summer 2007
 - Spring 2006
 - Summer 2006
 - Fall 2005

Publications

Drysdale, M.J., White, K.R., & Taylor, M.J. (Under Review). Boldfaced terms and their effects on student learning. *Teaching of Psychology*

Bates, S.C., Domenech Rodríguez, M., & Drysdale, M.J. (2007). Supporting and encouraging behavioral research among distance education students. *CUR Quarterly*, 28(1), 18-22.

Presentations

Drysdale, M.J. (2010). Boldfaced terms and their effects on student learning. The Intermountain Graduate Research Symposium, Logan, UT.

- Bigelow, J., Bates, S., & Drysdale, M. (2008). Introductory psychology textbook considerations: Incorporating qualitative and quantitative from student reviews. Rocky Mountain Psychological Association Convention, Boise, ID.
- Bates, S.C., Drysdale, M.J., Nelson, J.K., & Bigelow, J.C. (January, 2008). Top-down, bottom-up: using multiple data sources from multiple stake-holders in selecting an introductory psychology textbook. National Institute on the Teaching of Psychology, St. Petersburg Beach, FL.
- Hyte, H., Harward, R., Jensen, N., Drysdale, M., & White, K. (2006). Coordination among agencies to improve EHDI follow-up. National Early Hearing Detection and Intervention (EHDI) Conference, New Orleans, LA.
- Lubke, M. & Drysdale, M. (2006). Using developmental rate analyses to measure child outcomes. Council for Exceptional Children Convention and Expo, Salt Lake City, UT.

Teaching Assistantships

- Introductory Psychology (PSY 1010):
 - Fall 2006 – Spring 2007
 - Ran Laboratory section of course: developed & graded labs, quizzes, & exams; Lectured
- Introduction to Educational/Psychology Research (PSY 6570)
 - Fall 2005 – Summer 2006
 - Developed and graded quizzes and exams; Tutored students

Research Assistantships

- Multisensory Cognition Laboratory
 - Summer 2008 – Spring 2009
 - Managed and scheduled lab activities, including overseeing undergraduate experiences and evaluating their semester-long performance; Collected data with infants, young children, and college students; Ran experiments involving young children and their number abilities
- National Center for Hearing Assessment & Management
 - Fall 2005 – Summer 2006
 - Collected data on newborn hearing screening and children having received services at School for Deaf; Completed an evaluation of the effectiveness and communication of the two programs
- Summit Research Laboratory:
 - Summer 2004 – Summer 2005
 - Collected data on Safe & Drug Free Schools; Collaborated with Early Childhood educators in developing Utah Early Childhood Standards

Research Projects

- New Faculty Teaching Academy Evaluation
 - Fall 2008 – Spring 2009
 - Designed and implemented an evaluation of the New Faculty Teaching Academy at Utah State concerned with helping new faculty members improve their teaching abilities. This included developing evaluation questions, designing surveys, collecting data, comparing the Teaching Academy structure with similar programs from other universities, and writing a final report.

Mentoring Experience

- Undergraduate student mentoring
 - Spring 2010
 - Mentored an undergraduate student (psychology major) through the process of reviewing the literature, writing articles, and collecting and analyzing data.

International Experience

- Completed 2 years of volunteer work in Venezuela (April 1999 – April 2001); learned the Spanish language

Computer Literacy/Software Programs

Blackboard; SPSS; Microsoft Word, PowerPoint, Excel; GIMP; Inkscape; Blender; Kompozer

Professional Affiliations

- Psi Chi

References

Karl R. White, PhD
Professor of Psychology, Psychology
Department
Utah State University
Logan, UT 84322
(435) 797-3013

Dr. White served as my committee chair for my master's thesis and is serving as committee chair for my doctoral dissertation. He has also been my supervisor in a teaching assistantship and my mentor in conducting an evaluation of services provided by the Utah School for the Deaf.

J. Nicholls Eastmond, PhD
Professor, Instructional Technology
Department
Utah State University
Logan, UT 84322
(435) 797-2642

Dr. Eastmond supervised me in gaining experience with qualitative and ethnographic research techniques. I have also taken a course from Dr. Eastmond.

Byron R. Burnham, PhD
Vice Provost & Dean of Graduate
Schools

Dr. Burnham mentored me through writing a grant and has also supervised my efforts in evaluating the university's New Faculty

Utah State University
Logan, UT 84322
(435) 797-1193

Teaching Academy.