University of Louisville

ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

5-2012

Predictors of student course evaluations.

Timothy Michael Sauer
University of Louisville

Follow this and additional works at: <https://ir.library.louisville.edu/etd>

Recommended Citation

Sauer, Timothy Michael, "Predictors of student course evaluations." (2012). *Electronic Theses and Dissertations*. Paper 1266.
<https://doi.org/10.18297/etd/1266>

This Doctoral Dissertation is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

PREDICTORS OF STUDENT COURSE EVALUATIONS

By

Timothy Michael Sauer
University of Louisville

A Dissertation
Submitted to the Faculty of the
College of Education and Human Development
of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

College of Education and Human Development
University of Louisville
Louisville, Kentucky

May 2012

Copyright 2012 by Timothy Michael Sauer

All rights reserved

PREDICTORS OF STUDENT COURSE EVALUATIONS

By

Timothy Michael Sauer
B.A., Bellarmine University, 2006

A Dissertation Approved on

April 23, 2012

By the following Dissertation Committee:

Dr. Namok Choi
Dissertation Chair

Robert S. Goldstein

Dr. Joseph Petrosko

Dr. Rod Githens

Dr. Jill L. Adelson

ACKNOWLEDGMENTS

When I began the process of writing my dissertation some 18 months ago I looked forward to putting to words my sincere thanks and an acknowledgement of those individuals that have contributed to the completion of my dissertation. It is because of the meaningful contributions of the great educators I have encountered that I found my calling in the field of education research.

First and foremost I want to thank my mentor and dissertation chair, Dr. Namok Choi, for her tireless work in providing me with invaluable feedback and direction and keeping me on track during my dissertation journey. Without our post-it note contracts that hung on your office wall I would probably still be trying to decide on a research topic. Your instruction in statistics, that I proudly rate a 5 out of 5, provided me with a strong foundation that made me the researcher I am today. The first class I took as a graduate student was a statistics course taught by Dr. Choi. During the very first class meeting, in the summer of 2007, I immediately knew that I had chosen the right University at which to complete my post-baccalaureate studies.

In addition to your role as an educator and mentor you have been a wonderful friend. Your door is always open and there has always been a chair for me to sit in and enjoy a great conversation. From the bottom of my heart I say 감사합니다.

To the members of my dissertation committee, Dr. Jill Adelson, Dr. Rod Githens, Bob Goldstein, and Dr. Joe Petrosko, whose time and expertise were my greatest resource, I am forever indebted. Jill taught me the advanced statistical methods that

allowed me to undertake this study and my experience as her graduate research assistant gave me confidence in my ability to conduct such research. Jill is a true superwoman, providing me with answers to questions about HLM equations and multiple imputation methods at 2 AM, all while pregnant with her first child.

To Rod, my time as your graduate assistant was a tremendously rewarding experience but even more satisfying were the countless hours spent in your office conversing. It was during one of those conversations with Rod that the idea to explore student course evaluations was born.

To Bob, I want to thank you for your support and for believing in the merit of this research, without which I would not have been able to conduct my study. I must also thank you and your research staff for dedicating innumerable hours to procuring and de-identifying data. Becky Patterson and Arnold Hook were tremendous assets and must be commended for their work in linking and mining the data that were analyzed in my study.

To Joe, thank you for believing in me when I came to you as a student in counseling psychology and asked for admission into the ELFH program. Your extensive knowledge of student course evaluations, in particular the instrument used in the current study, was a great resource.

Additional thanks to the following individuals at the University of Louisville for their meaningful contributions to my doctoral journey: Dr. Linda Shapiro, Dr. Nancy Cunningham, and Kelly Ising.

ABSTRACT

PREDICTORS OF STUDENT COURSE EVALUATIONS

Timothy Michael Sauer

April 23, 2012

This dissertation explored the relationship between student, course, and instructor-level variables and student course ratings. The selection of predictor variables was based on a thorough review of the extensive body of existing literature on student course evaluations, spanning from the 1920's to the present day. The sample of student course ratings examined in this study came from the entirety of student course evaluations collected during the fall 2010 and spring 2011 semesters at the College of Education and Human Development at a large metropolitan university in the southern United States. The student course evaluation instrument is composed of 19 statements concerning the instructor's teaching ability, preparation, grading, the course text and organization to which the student rates their agreement with the statement on a 5 point Likert-type scale ranging from 1 "Strongly Disagree", "Poor", or "Very Low" to 5 "Strongly Agree", "Excellent" or "Very High".

In order to assess the relationship between the student, course, and instructor-level variables and the student course rating, hierarchical linear modeling (HLM) analyses were conducted. Most of the variability in student course rating was estimated at the student-level and this was reflected in the fact that most of the statistically significant relationships were found at the student-level. Prior student course interest and the amount

of student effort were statistically significant predictors of student course rating in all of the regression models. These findings were supported by previous studies and provide further evidence of such relationships.

Additional HLM analyses were conducted to assess the relationship between student course rating and final course grade. Results of the HLM analyses indicated that student course rating was a statistically significant predictor of student course grade. This finding is consistent with the existing literature which posits a weak positive relationship between expected course grade and student course rating.

TABLE OF CONTENTS

	PAGE
ACKNOWLEDGMENTS.....	.iii
ABSTRACT.....	v
LIST OF TABLES.....	x
LIST OF FIGURES.....	xi
CHAPTER	
I. INTRODUCTION.....	1
Statement of the problem.....	3
Research questions.....	3
Significance.....	4
Limitations.....	4
Definitions.....	4
II. LITERATURE REVIEW.....	6
History of student evaluations.....	6
Defining teacher effectiveness.....	9
Multidimensionality.....	11
Evaluation instrumentation.....	15
Online course evaluations.....	16
Reliability and validity of student evaluations.....	20
Reliability.....	20
Internal consistency reliability.....	20
Inter-rater reliability.....	20
Test-retest reliability (stability).....	21
Validity.....	23
Criterion validity.....	23
Construct validity.....	24
Predictive student evaluation variables.....	26
Administration of evaluations.....	26
Timing of evaluation.....	27
Anonymity of student raters.....	28
Instructor presence in classroom.....	28
Course characteristics.....	29
Electivity.....	29
Class meeting time/length.....	31
Class size.....	32
Workload/rigor.....	33
Instructor characteristics.....	34
Instructor experience.....	36
Gender.....	38
Ethnicity.....	39

	Student characteristics.....	40
	Prior interest in subject.....	40
	Gender.....	41
	Age.....	42
	Expected grade.....	43
	Summary.....	45
	Research questions.....	48
III.	METHODS.....	50
	Participants.....	50
	Participant demographics.....	51
	Instruments.....	53
	Student course evaluation.....	53
	Data source.....	54
	Variables.....	55
	Procedure.....	57
	Analysis.....	58
	Research question one.....	58
	Research question two.....	59
	Fully unconditional model.....	60
	Random coefficients model.....	61
	Contextual model.....	63
	Class level.....	63
	Instructor level.....	64
	Research question three.....	65
	Fully unconditional model.....	66
	Contextual model.....	66
IV.	RESULTS.....	68
	Descriptive Statistics.....	68
	Research Question One.....	71
	Construct validity.....	71
	Reliability.....	76
	Research Question Two.....	78
	Overall student course rating.....	78
	Unconditional model.....	78
	Student level model.....	79
	Class level model.....	79
	Instructor level model.....	79
	Factors as outcome variables.....	80
	Estimated variance.....	81
	Factor one.....	81
	Factor two.....	82
	Factor three.....	82
	Summary.....	83
	Research Question Three.....	84
	Unconditional model.....	84
	Control model.....	85

	Student rating as a predictor model.....	85
V.	DISCUSSION.....	87
	Review of the Results.....	87
	Research question one.....	87
	Research question two.....	88
	Research question three.....	89
	Emergent Factors.....	89
	Predictors of Student Course Rating.....	90
	Student-level predictors.....	90
	Course-level predictors.....	93
	Instructor-level predictors.....	95
	Course Rating as a Predictor of Final Course Grade.....	96
	Limitations.....	97
	Summary and recommendations.....	99
	REFERENCES.....	102
	APPENDICES.....	116
	CURRICULUM VITAE.....	155

LIST OF TABLES

TABLE	PAGE
1. Comparison of student evaluation factor models.....	14
2. Summary of the variables influencing student ratings.....	46
3. Sample demographics: Means and standard deviations.....	52
4. Sample demographics: Frequencies for binary variables.....	53
5. Descriptive statistics for the predictor variables.....	69
6. Descriptive statistics for the outcome variables.....	70
7. Pattern Coefficients, Structure Coefficients, and Communalities.....	74
8. Reliability statistics for obtained scores.....	77
9. Estimated variance at the student, course, and instructor level.....	81
10. Summary of the statistically significant predictors of student course rating...	84

LIST OF FIGURES

FIGURES	PAGE
1. Prediction model.....	49

CHAPTER I

INTRODUCTION

One of the most commonly used indicators of instructor performance in higher education is the student course evaluation. The resultant student data are often one of the only sources of information pertaining to the instructor's teaching effectiveness and at many postsecondary institutions the student data are relied upon by administrators in making personnel and tenure decisions (Braskamp & Ory, 1994; Centra, 1993; Wachtel, 1998).

Rating scales are the most commonly used type of student course evaluation instrument. Rating scale instruments contain items with a limited range of responses, usually between three and seven response options on a continuum from "strongly agree" to "strongly disagree" or "very important" to "not at all important" (Braskamp & Ory, 1994). In 1999, nearly 90% of 600 liberal arts colleges surveyed reported the use of student rating scales (Seldin, 1999). This number has grown substantially within this sample of 600 liberal arts colleges over the past several decades, from 67.5% in 1983 to 80.3% in 1988 to 86% in 1993 (Seldin, 1993). The proportion of large research universities using student rating scales of teacher effectiveness has been estimated as high as 100% (Ory & Parker, 1989).

Given the prevalence of their use in postsecondary institutions, there exists a substantial body of literature on student evaluations of instructor effectiveness. Current estimates of the amount of published research on student evaluation of instructor

effectiveness range from 1,300 (Cashin, 1995) to 2,000 (Feldman, 2003) citations. The research is so vast that periodically reviews of existing literature are published. Feldman (2003) cites 34 such reviews, having authored 15 meta-analyses of the relevant literature himself. Within this paper, the citations range in publication date from 1928 to 2010, covering a span of over 80 years.

While many researchers contend that scores obtained from current student evaluation instruments are valid and reliable measures of instructor effectiveness (Aleamoni, 1999; Cashin, 1995; Centra, 1993; Costin, Greenough, & Menges, 1971; Firth, 1979 Marsh, 1984; Marsh & Overall, 1980), there is still a large contingent that argue that the results from such instruments should not be relied upon for making personnel and tenure decisions (Wachtel, 1998). Opponents of the use of student evaluation of instructor effectiveness cite several concerns: (a) there is no consensus definition of effective teaching, (b) teaching to promote positive evaluations may conflict with good teaching practice, and (c) evaluation scores may be influenced by variables (biases) that have nothing to do with instructor effectiveness (Wachtel, 1998).

Centra (1993) defines bias in this context as “a circumstance that unduly influences a teachers’ ratings, although it has nothing to do with the teacher’s effectiveness” (p.65). He argues that most individual student, course, or teacher characteristics do not have an undue influence but in combination may. When student evaluations are collected for self-improvement purposes, biases can be addressed by collecting additional data or dismissing the results. When used for personnel decisions, it is important that any possible bias to student evaluations is empirically studied and controlled for (Braskamp, Brandenburg, & Ory, 1984; Centra, 1993).

The majority of the published literature on student ratings of instructor effectiveness is focused on the exploration of the potential student, course, and instructor-level biases. Despite the abundance of empirical research on the relationship between these potential biasing characteristics and student ratings, there remains a great deal of uncertainty about the true nature of these relationships. Contradictory results are a common thread in much of the student evaluation literature, resulting in inconclusive evidence of the presence or absence of such a bias. This simply reinforces the need for future research in the area and provides justification for the current study.

Statement of the Problem

The nature of the relationship between many of the potential biasing variables and student ratings of instructor effectiveness remains inconclusive. This is a result of both contradictory findings (e.g. student gender, instructor gender, timing of evaluation, and course workload) and limited published literature (e.g. instructor ethnicity and class meeting time). Because of the high stakes personnel and tenure decisions made in part based upon student ratings data, it is of the utmost importance to accurately assess the potential biasing effect of student, course, and instructor-level variables. The purpose of this study is to assess the effect of the potential student, class, and instructor-level biasing variables on student ratings of instructor effectiveness.

Research Questions

The research questions being addressed in this study are as follows:

1. Do the student evaluation ratings obtained in the study exhibit adequate reliability and construct validity?
2. How are student, course, and instructor-level variables related to student ratings of instructor effectiveness?
3. Do a students' ratings of instructor effectiveness predict students' final course grades?

Significance

The major contribution of this study is its addition to the body of literature on the relationship between student, course, and instructor-level variables and student ratings of instructor effectiveness. While the findings may not be universally generalizable, the results can be considered as additional data to be considered in assessing the effect of biasing variables. Future researchers and meta analysts can consider the findings as additional evidence in coming to a consensus decision about the impact of the student, course, and instructor-level variables. Additionally, this study incorporates several variables that have not been widely used in previous research (e.g. instructor ethnicity and class meeting time).

Limitations

As stated previously, the results of this study cannot be generalized to the worldwide population of college students. The sample in the study is limited to undergraduate and graduate students enrolled in a college of education and human development at a large metropolitan research university in the southern United States. Another limitation is the prevalence of missing data. Given the fact that data were merged from several different university maintained databases, there was some missing information. These missing data occurred across the student, class, and instructor level. Perhaps the most concerning limitation is the fact that there are a large number of students that did not complete the optional course evaluation. In the current study the mean response rate for the sampled courses was 55.9% with individual course response rates ranging from 7% to 100%. Without having data from the non-respondents it is unclear how the participant and non-participant students may have differed in their assessment of the course and instructor.

Definitions

Operational definitions for all of the variables included in the study are provided in the variable section of Chapter 3 (p.53-55). Within the context of this study the following definitions were used.

Instructor effectiveness.

Instructor effectiveness is defined as producing “beneficial and purposeful student learning through the use of appropriate procedures” (Centra, 1993; p. 42). These procedures include what the instructor does to organize and run the course, and account for the classroom atmosphere, learning activities, method of content delivery, workload and assignments (Braskamp & Ory, 1994). The term instructor effectiveness is used interchangeably with teacher effectiveness.

Student evaluation of instructor effectiveness.

Student evaluation of instructor effectiveness (SET) is defined as an instrument completed by students enrolled in a course to assess student perceptions of the instructor’s ability to facilitate learning. This broad term encompasses various instruments of differing delivery methods. The term student evaluation of instructor effectiveness is used interchangeably with student ratings, student ratings of instructor effectiveness, student evaluation of teacher effectiveness, and student ratings of teacher effectiveness.

In the body of this paper class-level and course-level both refer to the same level of specification and are used interchangeably.

CHAPTER II

LITERATURE REVIEW

History of Student Evaluations

The evaluation of instructor effectiveness can be traced back to the universities of medieval Europe. A committee of students, selected by the rector, reported instances in which the instructor failed to adhere to the course schedule. These violations of the course schedule resulted in monetary fines that continued each day the professor remained off schedule (Centra, 1993, citing Rashdall, 1936).

Modern evaluation practices began in the early 1800's, when Boston schools were inspected by committees of local citizens to determine whether instructional goals were being met (Spencer & Flyr, 1992). In time, these "inspections" became in-house procedures mandated for all instructional personnel at educational institutions. The most commonly used instrument to record observations from these "inspections" was the teacher rating scale, the first of which appeared in the 1915 yearbook of the National Society for the Study of Education (Medley, 1987; Spencer & Flyr, 1992).

In the 1920s researchers began to explore the factors that may affect student evaluations of teacher effectiveness (Wachtel, 1998). One of the early pioneers in the field was Hermann Henry Remmer, who explored the relationship between student evaluations and course grades, the reliability of evaluation scores, and the similarities between student and alumni evaluation scores (Centra, 1993). In addition to his contributions to student evaluation research, Remmer and his colleagues at Purdue

University published the Purdue Rating Scale for Instructors (1927), considered to be the first formal student evaluation form (Centra, 1993). During this same period, formal student evaluation procedures were introduced at several other major United States universities (Marsh, 1987; Wachtel, 1998).

The Purdue Rating Scale is a graphic scale in which students rate an instructor on 10 qualities believed to be indicative of successful teaching: (a) interest in subject, (b) sympathetic attitude toward students, (c) fairness in grading, (d) liberal or progressive attitude, (e) presentation of subject matter, (f) sense of proportion and humor, (g) self-reliance and confidence, (h) personal peculiarities, (i) personal appearance, and (j) stimulating intellectual curiosity (Stalnaker & Remmers, 1928). A factor analysis of the Purdue Rating Scale indicated that the 10 items load on two unique teacher traits, an empathy trait and professional maturity trait (Smalzried & Remmers, 1943).

Student unrest and protest in the 1960s triggered a renewed interest in the use of student evaluations to assess instructor effectiveness. Unhappy with the quality of education, students demanded a voice in evaluating and improving their education. As a medium to express this voice, students administered, scored, and published their own evaluations of instructors. This haphazard system led the universities to intervene and develop and implement their own evaluation instruments (Centra, 1993).

Centra (1993) describes the 1970s as the golden age of research on student evaluation, during which studies were conducted that demonstrated the validity and reliability of student evaluation instruments and supported the utility of such instruments in academic settings (Wachtel, 1998). Modern-day research has continued to build upon previously published findings, employing advanced methods like meta-analysis and

hierarchical linear modeling. Other paths of research have investigated the feasibility of alternative methods of student evaluations, such as letters written by students and faculty developed narratives.

Current estimates of the amount of published research on student evaluation of instructor effectiveness range from 1,300 (Cashin, 1995) to 2,000 (Feldman, 2003) citations. The research is so vast that periodically reviews of existing literature are published. Feldman (2003) cites 34 such reviews, having authored 15 meta-analyses of the relevant literature himself. Within this dissertation, the citations range in publication date from 1928 to 2010, covering a span of over 80 years. Having said this, the purpose of this literature review is to provide an extensive overview of the existing literature on student evaluations of instructor effectiveness. For further reading on the published literature, readers are pointed towards the work of Feldman (1976a, 1976b, 1977, 1978, 1979, 1983, 1984, 1986, 1987, 1989a, 1989b, 1990, 1992, 1993, 2003), Centra (1993), Wachtel (1998), Cashin (1988, 1995), Marsh (1987), and Aleamoni (1999).

The majority of the published literature on student ratings of instructor effectiveness is focused on the exploration of the potential student, course and instructor-level biases. The body of literature has evolved in such a way that most studies build upon the empirical findings posited by previous authors, investigating the relationships between potential biasing variables and student ratings in different samples, with different evaluation instruments and differing sets of predictor variables. Because of the distinct nature of the student evaluation literature, this review is constructed within a similar framework, using previous empirical studies to create a prediction model.

Defining Teacher Effectiveness

There are no universally accepted criteria for assessing teacher effectiveness, but there are two factors common amongst many definitions: the outcome of student learning and the procedure. Centra (1993) accounts for both of these dimensions in his definition of effective teaching as producing “beneficial and purposeful student learning through the use of appropriate procedures” (p. 42). Braskamp, Brandenburg, and Ory (1984) echo this sentiment in describing the three major areas for defining effective teaching as input, process, and product. Input attempts to account for preexisting factors, such as student, teacher, and course characteristics that may influence the process and product. Process describes what the instructor does to organize and run the course, accounting for the classroom atmosphere, learning activities, method of content delivery, workload, and assignments. Product takes into account student learning outcomes. Braskamp, Brandenburg, and Ory argue that to fully evaluate instructor effectiveness all three aforementioned areas must be considered.

Following a similar structure of defining teacher effectiveness, outcome, and procedure, Fuhrmann and Grasha (1983) present three definitions of effective teaching based on the behaviorist, cognitive, and humanistic theories of learning. The behaviorist definition of effective teaching “is demonstrated when the instructor can write objectives relevant to the course content, specify classroom procedures and student behaviors needed to teach and learn such objectives, and show that students have achieved the objectives after exposure to the instruction” (Fuhrmann & Grasha, 1983, p. 287). The cognitive definition of effective teaching “is demonstrated when instructors use classroom procedures that are compatible with a student’s cognitive characteristics, can organize and present information to promote problem solving and original thinking on

issues, and can show the students are able to become more productive thinkers and problem solvers” (Fuhrmann & Grasha, 1983, pp. 287-288). The humanistic definition of effective teaching “is effective when teachers can demonstrate that students have acquired content that is relevant to their goals and needs, that they can appreciate and understand the thoughts of feelings of others better, and that they are able to recognize their feelings about the content” (Fuhrmann & Grasha, 1983, p. 288).

Feldman (1976b) synthesized the body of literature examining how college students define effective instruction. Forty-nine studies were identified and divided into two categories, structured response in which the student ranked a preset list of instructor characteristics and unstructured response in which the student responded freely with their own characteristics. In order to increase comparability among studies, student rankings were standardized (the individual ranking was divided by the total number of characteristics). The highest ranked characteristics for the structured response sample were instructor knowledge, stimulation of interest, class progress, and clarity of explanation. The highest ranked characteristics for the unstructured response sample were instructor concern and respect for students, instructor knowledge, stimulation of interest, and instructor availability or helpfulness. In a follow-up study, Feldman (1988) analyzed past studies ($n=18$) that had both teachers and students rank characteristics of effective instruction. The results indicate that students most valued teacher sensitivity and concern, organization of the course, teacher’s knowledge, and teacher’s stimulation of interest in the subject in defining effective instruction. Teachers ranked teacher’s knowledge, teacher’s enthusiasm, teacher’s sensitivity and concern, organization of

course, and clarity and understandableness as the most important indicators of effective instruction.

The results of Feldman's studies add further credence to the notion that there is not a singularly accepted definition of effective instruction, as evidenced by the variability in student and teacher responses. While there is variability in the definition, the results indicate that both students (structured and nonstructured respondents) and teachers view content knowledge, empathy, and clarity/organization as important indicators of effective instruction.

Multidimensionality

The lack of a clear definition of effective instruction may be indicative of different emphases placed on various aspects of effective teaching, or it may be due to the multidimensional nature of the construct (Patrick & Smart, 1998). Factor analytic studies have provided some support for the multidimensionality of teaching effectiveness, and as such, any evaluation of teaching performance should account for this multidimensionality (Abrami, d'Apollonia, & Cohen, 1990; Cashin, 1995; Marsh & Dunkin, 1992). The scaled global score often reported with evaluation instruments falls short of accounting for this multidimensionality, lacking the sophistication to provide feedback on specific instructor behaviors. The use of factor scores, composed of subsets of items, provides for a more meaningful interpretation of the findings and a reflection of the multidimensionality of the construct (Algozzine et al., 2003).

Multiple authors have proposed factor models for describing the construct of instructor effectiveness. Several of the more prominent models are outlined below. The Students' Evaluation of Educational Quality (SEEQ) instrument proposes a nine-factor

model of teacher effectiveness: (a) learning/value, (b) instructor enthusiasm, (c) organization/clarity, (d) group interaction, (e) individual rapport, (f) breadth of coverage, (g) examinations/grading, (h) assignments/grading, and (i) workload/difficulty (Marsh, 1983, 1984, 1987). These nine factors were developed based on a review of existing student evaluations of instructor effectiveness (SETs) and the relevant theories and literature, interviews with teachers and students, and psychometric analyses. This nine-factor model has been confirmed in more than 30 published studies (Marsh, 1987).

Centra (1993) and Braskamp and Ory (1994) propose a six-factor model of teaching effectiveness that is similar in content to Marsh's model, including several of the same factors and collapsing some of the factors in Marsh's model into single factors. The six factors include: (a) course organization and planning, (b) clarity/communication skills, (c) teacher-student interaction/rapport, (d) course difficulty/workload, (e) grading and examinations, and (f) student self-rated learning (Cashin, 1995; Centra, 1993; Braskamp & Ory, 1994).

Patrick and Smart (1998) conducted a two-phase study to develop a model for understanding effective instruction and an instrument for measuring this construct. In the first phase, 148 undergraduate students completed a qualitative questionnaire that asked them to record in their own words the attributes, qualities, and characteristics of an effective teacher. The qualitative data were analyzed and categorized into 36 thematic groups of teacher attributes. The resultant 36 attributes from the qualitative phase were combined with items from existing widely used measures of instructor effectiveness to create a 72-item 5-point Likert scale (from 1 "does not describe teacher very well at all" to 5 "describes the teacher perfectly") meta-inventory. Two hundred and sixty-six

undergraduate psychology students completed the meta-inventory, which asked them to respond to the 72 statements while thinking of a teacher from any point of their education that they found to be the most effective.

A principal components factor analysis revealed a 24-item three-factor solution (student respect, organization and presentation skills, and ability to challenge students), which accounted for 44.1% of the total variance. Each of the three factors were composed of eight items and exhibited acceptable internal reliability estimates of .86 (student respect), .83 (organization and presentation skills) and .79 (ability to challenge students). Patrick and Smart (1998) provided additional evidence of the plausibility of the three-factor solution by comparing their model to the work of other scholars. Aligning closest with Patrick and Smart's model of effective instruction was Brown and Atkins' (1988) three-factor model of caring, systematic, and stimulating.

The models in the above section provide empirical evidence of the multidimensionality of the instructor effectiveness construct (Table 1). There are commonalities among the models. The factor of organization and presentation skills (Patrick & Smart, 1998) is similar to Marsh's (1987) organization/clarity factor and Centra (1993) and Braskamp and Ory's (1994) course organization and planning and clarity, communication skills factors. The student respect factor (Patrick & Smart, 1998) is comparable to Marsh's (1987) group interaction and individual rapport factors as well as Centra (1993) and Braskamp and Ory's (1994) teacher student interaction/rapport factor. Patrick and Smart's (1998) ability to challenge students factor can be compared to Marsh's (1987) workload/difficulty, examinations/grading and assignments/grading factors as well as Centra's and Braskamp and Ory's course workload/difficulty and

grading and examination factors. While there is some variability amongst the models, this may be due in part to the fact that in each study a different instrument was analyzed. These instruments may have varied in how they emphasized the different aspects of instructor effectiveness. Additionally, there are an infinite number of rotations possible for any set of data. The factors and their definitions depend on the interpretation of the individual researcher (Patrick & Smart, 1998).

Table 1

Comparison of student evaluation factor models

Author(s)	Number of factors	Factors
Marsh (1983, 1984, 1987)	9	learning/value, instructor enthusiasm, organization/clarity, group interaction, individual rapport, breadth of coverage, examinations/grading, assignments/grading, workload/difficulty
Centra (1993); Braskamp and Ory (1994)	6	course organization and planning, clarity and communication skills, teacher-student interaction/rapport, course difficulty/workload, grading and examinations, student self-rated learning
Patrick and Smart (1998)	3	student respect, organization and presentation skills, ability to challenge students

Evaluation Instrumentation

Rating scales are the most commonly used student evaluation instruments. In 1999, nearly 90% of 600 liberal arts colleges surveyed reported the use of student rating scales (Seldin, 1999). This number has grown substantially within this sample of 600 liberal arts colleges over the past several decades, from 67.5% in 1983 to 80.3% in 1988 to 86% in 1993 (Seldin, 1993). The proportion of large research universities using student rating scales of teacher effectiveness has been estimated at 100% (Ory & Parker, 1989).

Rating scale instruments contain items with a limited range of responses, usually between three and seven response options on a continuum from “strongly agree” to “strongly disagree” or “very important” to “not at all important” (Braskamp & Ory, 1994). Three common types of rating scales are the (a) omnibus form, (b) goal-based form and (c) form based on the cafeteria system. An omnibus instrument is standardized, contains a fixed set of items, and is administered to students in all classes across multiple departments and schools, allowing for comparisons across faculty. The instruments are often statistically divided into subscales of the larger instructor effectiveness construct. A goal based form has students rate their own performance on stated course goals and objectives (e.g. gaining knowledge of the subject, developing skills, or gaining appreciation of subject) instead of assessing the performance of the instructor (Braskamp & Ory, 1994).

Prior to the development of the cafeteria system at Purdue University in the 1970’s, campus-wide evaluation instruments included the same items for every professor. The cafeteria system introduced a bank of items from which individual faculty or an academic department can select the items that are aligned closest with the objectives and

goals of the course(s). Most cafeteria systems include a set of global items that are common across all evaluations and used to summarize the students overall evaluation of the instructor's effectiveness. These may include items such as: "overall this is an excellent course" or "overall the instructor is an excellent teacher (Braskamp & Ory, 1994)."

Online course evaluations. A more recent development in the administration of student course evaluations has been online delivery. Hmielecki (2000) surveyed 200 of the most wired colleges in the United States and found that only two were using online evaluation systems but nearly 25% reported that they planned to move to online evaluations in the future. While online course evaluations are not the most prevalent method of administration, there is clear evidence of expected growth in its usage.

Proponents cite several advantages to the use of online course evaluations, including: (a) the lower cost of online evaluations in comparison to traditional paper-and-pencil evaluations, (b) online evaluations require less class time, (c) online evaluations are a "greener" alternative to the paper-heavy traditional evaluations, (d) online evaluations allow for instantaneous feedback because there is no additional data input required, (e) students may feel greater anonymity due to the removal of any hand-written components, and (f) students are free to complete online evaluations at their convenience (Anderson, Cain, & Bird, 2005; Johnson, 2002). Questions remain about the response rate for online evaluations and how student responses may differ when collected online compared traditional paper-and-pencil administration.

Because of the emerging nature of online course evaluations, there is a limited amount of published empirical research on the effect of evaluation delivery method on

response rate. Layne, Decristoforor, and McGinty (1999) compared online to traditional evaluation scores in a sample of 66 classes and reported a response rate of 47.8% for the online group and 60.6% for the traditional group. Johnson (2002) conducted several pilot tests prior to the implementation of an online evaluation system at Brigham Young University. In 1997, 36 courses were evaluated online, yielding a response rate of 40%. In 1999, 194 courses were evaluated online, yielding a response rate of 51%. The final pilot test involved the online evaluation of 47 courses with 3,076 students. This yielded a response rate of 62% (Johnson, 2002).

Thirty-four of the participating faculty in the Johnson's (2002) third pilot test reported the nature of their communication with students regarding the evaluation and the corresponding response rate. Faculty members that assigned students to complete the online evaluation and awarded bonus points for doing so achieved the highest mean response rate at 87%, with a range from 59 to 95%. Faculty members that assigned students to complete online evaluations but not awarding points achieved a mean response rate of 77%. Faculty that encouraged students to complete the online evaluation but without making it a formal assignment achieved a mean response rate of 32%. The lowest mean response rate at 20% came from faculty members that did not mention the evaluation form to students (Johnson, 2002).

Dommeyer, Baum, Chapman, and Hanna (2003) compared the response rates for paper-and-pencil to online evaluations within a sample of classes taught by 16 business school professors. Response rates were lower (29%) in the online format than the traditional paper-and-pencil method (70%). When any type of grade incentive (reporting

grades early or a .04% increase in grade for completing the evaluation) was used, the online format was comparable to the traditional methods.

In contrast to the above findings are the results reported by Anderson, Cain, and Bird (2005). An online course evaluation was piloted in a sample of three courses in the College of Pharmacy at the University of Kentucky. The online evaluation format yielded response rates of 85%, 89%, and 75% in the respective courses. These response rates were slightly higher than the traditional paper-and-pencil format rate of 80%. Other divergent evidence comes from Chang (2004), who reported response rates of 79% for paper-and-pencil evaluations and 95.3% for online evaluations in a sample of 1,052 courses.

The limited published results suggest that online student course evaluations may achieve lower response rates than the traditional paper-and-pencil format. There is limited evidence that the response rate for online evaluations may be higher if students are presented with incentives to complete the evaluation. The literature suggests several strategies for increasing the response rate, including: (a) instructors encouraging students to complete the evaluations, (b) providing an explanation of what the evaluation results are used for, (c) granting early access to grades for completing the evaluation, (d) providing bonus points for completing the evaluation, (e) early access to registration for evaluation completers, or (f) the use of prizes that can be won by evaluation completers (Anderson et al., 2005; Chang, 2004; Johnson, 2002).

Based on a limited number of empirical studies, there does not appear to be a consensus opinion on the relationship between evaluation delivery method and student evaluation scores. In a sample of 74 courses that were administered both online and

paper-and-pencil evaluations, Johnson (2002) found a correlation of .86 on the overall course evaluation items between the two delivery methods. The online overall course evaluations were on average .01 points higher than the paper-and-pencil scores. The author did not report the results of any statistical tests of the difference in means. Layne, Decristoforor, and McGinty (1999) compared online to traditional evaluation scores in a sample of 66 classes and did not find a statistically significant difference in the means. Paolo, Bonaminio, Gibson, Partridge, and Kallail (2000) compared online to mailed student course ratings of fourth-year medical students and reported that there were no statistically significant differences between the two groups on any of the 62 items.

Chang (2004) compared paper-and-pencil to online course evaluation results in a sample of 624 undergraduate courses at a teachers' college in Taiwan. Class sizes ranged from 5 to 51 students. Results indicated that paper responses were statistically significantly ($p < .001$) higher than the online responses for each of the 13 items of the course evaluation instrument. Additionally, t-test results indicate that the scores for each of the four factors that compose the evaluation form as well as the summative measure of overall course evaluation were significantly higher for the paper responses. The author attributes this difference to the lower degree of anonymity in the paper-and-pencil setting. It should be noted that the student participants were informed that the purpose of the study was to compare evaluation scores for the online and paper-and-pencil format.

The studies above report conflicting results about the difference between online and paper-and-pencil course evaluations. In order to make more conclusive statements about the relationship, there is a need for further study in the area.

Reliability and Validity of Student Evaluations

Reliability

Internal consistency reliability. Research provides evidence of high internal consistency of the scores obtained from various student evaluation instruments, with several authors reporting coefficients in the .90 range (Aleamoni, 1999; Centra, 1993; Marsh, 1984). Internal consistency can be defined as the degree to which items on an instrument measure that attribute in a consistent manner (Tashakkori & Teddlie, 1998). It is determined by calculating the average correlation between items on the instrument. Marsh (1984) cautions that internal consistency coefficients provide an inflated estimate of the reliability of student evaluations because it ignores the error due to lack of agreement amongst students.

VanLeeuwen, Dormody, and Seevers (1999) presented generalizability theory analysis as one alternative method of assessing the reliability of SETs because of its ability to accurately partition variance amongst classes, items and students. By averaging each student's response to all items, VanLeeuwen et al (1999) obtained a reliability estimate of .957, slightly lower than the Cronbach's alpha of .97. Averaging over both items and students within a class, reliability estimates ranged from .80 in a class with seven students to .96 in a class with 47 students.

Inter-rater reliability. There is evidence of sufficient inter-rater reliability of scores obtained through student evaluation instruments. Inter-rater reliability can be defined as the degree to which ratings by two or more raters are consistent with one another (Tashakkori & Teddlie, 1998). A commonly used method of assessing the extent of agreement within a class of students is the computation of the intraclass correlation

coefficient (Centra, 1993). These correlations should be interpreted with caution as they are highly influenced by the number of raters. The correlation between any two students in the same class is generally low, in the .20s (Centra, 1993; Marsh, 1984). As the number of raters increases, the reliability coefficient (intraclass correlation) increases.

Marsh (1984) found that the inter-rater reliability for the Students' Evaluations of Educational Quality (SEEQ) factors to be about .23 for one student, .60 for five students, .74 for ten students, .90 for 25 students, and .95 for 50 students in the same class. Centra (1993) calculated the reliability for the overall teacher rating on the Student Instructional Report (SIR) and found coefficients similar to those reported by Marsh; .65 for five students, .78 for 10 students, .90 for 25 students, and .95 for 50 students. Cashin (1995) reported slightly lower reliability coefficients for the IDEA Overall Evaluation, with median reliabilities of .69 for 10 students, .83 for 15 students, .83 for 20 students, .88 for 30 students, and .91 for 40 students. One note of caution is that measures of inter-rater reliability may provide an inflated/deflated estimate of the reliability of student evaluations because of the influence of the number of raters.

Test-retest reliability (stability)

Several studies have been published that explore the stability of evaluation scores over time. Test-retest reliability or stability may be defined as the degree to which repeated administrations of a test differentiate members of a group in a consistent manner, determined by calculating the correlation between two administrations of the instrument in the same group of individuals (Tashakkori & Teddlie, 1998). The consensus amongst the literature is that ratings of the same instructor by the same students tend to be stable over time (Cashin, 1995; Costin, Greenough, & Menges, 1971;

Marsh, 1984). In one of the earliest studies of the stability of student evaluation scores, Guthrie (1954) reported correlations of .87 and .89 between student's evaluation scores for an instructor from one year to the next. Costin (1968) compared student's mid-semester and end-of-semester ratings and found moderate to high correlations on the four measured factors of instructor effectiveness (.70-.87).

There are some practitioners that question the ability of students to recognize effective teaching while they are enrolled in the course. These individuals argue that a student can not accurately assess the effectiveness of the instructor until they are called upon to utilize the course content in a real-life situation or later coursework (Marsh, 1984). In an attempt to account for the real-life utilization of skills taught in college courses, the following studies compared retrospective scores to scores obtained after graduation. Marsh and Overall (1980) conducted a longitudinal study of over 100 college courses, comparing student ratings of teacher effectiveness at the end of the semester and at an additional time point several years following the course, at least one year after the student's graduation. The researchers reported a correlation of .83 between the end of semester and later evaluation scores. Firth (1979) correlated course ratings obtained at graduation and one year after graduation and reported findings similar to Marsh and Overall (1980).

As illustrated by the examples above, student ratings of teacher effectiveness tend to exhibit stability over time. Additionally, the effect of real-life experience and the utilization of course knowledge have minimal impact on a student's rating of instructor effectiveness. Student ratings taken at the time of the course do not significantly change over time.

Validity

The underlying question in assessing the validity of an instrument is whether the instrument measures what it is supposed to measure. In the case of student evaluations, the expected outcome is a measure of the course instructor's effectiveness. Given that there is no consensus definition of instructor effectiveness or a predominant agreement on the number of dimensions underlying the construct, it is difficult to assess whether a student evaluation instrument measures the construct of instructor effectiveness (Cashin, 1988; Marsh, 1984). Nevertheless, some researchers attempted to establish evidence for validity of the scores generated from several measures of teaching effectiveness.

Criterion validity. Criterion validity is assessed by examining the correlation between the instrument under investigation and a criterion variable that is representative of the construct. One of the most widely used criteria in assessing the criterion validity of instructor effectiveness is a measure of student learning (Cashin, 1988; Marsh, 1984). Because of the variability across individual course examinations and the often subjective nature of such assessments, it is difficult to assess the relationship between evaluation scores and student learning. This type of investigation may be possible in large multisection courses with standardized course content and examinations taught by different professors (Marsh, 1984; Marsh & Roche, 1997).

Results of multisection validity studies have demonstrated that classes with the highest evaluation ratings also have the highest levels of student learning as measured by scores on course examinations (Marsh & Roche, 1997). Cohen (1987) conducted a meta-analysis of 41 multisection validity studies and found that mean correlations between final course examinations and the student evaluation subscales were .55 for structure, .52

for interaction, .50 for skill, .49 for overall course, .45 for overall instructor, .39 for learning, .32 for rapport, .30 for evaluation, .28 for feedback, .15 for motivation, and -.04 for difficulty. All of the correlations aside from the motivation and difficulty subscales were statistically significant. d' Appolonia and Abrami (1997) utilized a similar method in analyzing the results of 43 multisection validity studies and reported a mean correlation coefficient between student evaluation and student learning of .47 with a 95% confidence interval between .43 and .51. The results of both of the reported meta-analyses are indicative of a moderate to large association between student evaluation ratings and student learning outcomes, thus providing evidence of criterion validity.

Construct validity. Construct validity, meaning that the instrument highly correlates with similar measures of the construct under investigation (convergent validity) and correlates less with dissimilar measures of the construct (discriminant validity) have been assessed by conducting multitrait-multimethod studies (Greenwald, 1997).

Marsh (1982) administered a student evaluation instrument to students and faculty (self-rated) in 329 different classes at the University of Southern California. Results of a factor analysis revealed nine separate evaluation traits or factors underlying the evaluation instrument: (a) learning/value, (b) instructor enthusiasm, (c) organization, (d) group interaction, (e) individual rapport, (f) breadth of coverage, (g) examinations, (h) assignments, and (i) workload/difficulty. This nine-factor structure was upheld for both the student and faculty responses.

Marsh (1982) employed the Campbell-Fiske analysis (1959) to assess the convergent and discriminant validity. The multiple traits were the nine evaluation factors presented in the paragraph above, and the multiple methods were the distinct groups of

ratars: students and faculty members. The convergent validity was evaluated by correlating the same traits across the student and faculty ratings, with a median r of .45 and statistically significant correlations for all of the traits. This finding is consistent with previous results published by Doyle and Crichton (1978, $r = .47$), Webb and Nolan (1955, $r = .62$), and Marsh, Overall, and Kesler (1979, $r = .49$). Discriminant validity was assessed by (a) examining whether student-faculty agreement on each factor is independent of agreement on the other factors, and (b) looking for a method/halo effect as a source of method variance. Results of the Campbell-Fiske analysis provide evidence of both convergent and discriminant validity.

Further evidence of the convergent validity of student evaluation instruments has been demonstrated by comparing student evaluation ratings with those of trained observers (Centra, 1993; Marsh & Roche, 1997). Similar to the examples above that compared student ratings to self-ratings by faculty, these studies hypothesized that student evaluation ratings would correlate positively with the assessment of trained observers. Murray (1983) trained external observers to report on teacher behaviors collected during three one-hour periods. Observational data were collected for each of 54 college instructors by six to eight trained observers, totaling between 18 and 24 hours of observation per instructor. Results of the study indicated that instructors that had been rated highly by student evaluations exhibited behaviors consistent with effective teaching. The researchers concluded that teachers with high ratings taught differently than teachers with average or poor student evaluation ratings.

While many researchers will contend that scores obtained from current student evaluation instruments are valid and reliable, there is still a large contingent that will

argue that the results from such instruments should not be relied upon for making personnel and tenure decisions (Wachtel, 1998). Opponents of the use of student evaluation of instructor effectiveness cite several concerns: (a) there is no consensus definition of effective teaching, (b) teaching to promote positive evaluations may conflict with good teaching practice, and (c) evaluation scores may be influenced by variables (biases) that have nothing to do with instructor effectiveness (Wachtel, 1998). In the predictive student evaluation variables section, the literature pertaining to possible biases to student evaluations is reviewed in detail.

Centra (1993) defines bias as “a circumstance that unduly influences a teachers’ ratings, although it has nothing to do with the teacher’s effectiveness (p.65).” He argues that most individual student, course, or teacher characteristics do not have an undue influence but may in combination. When student evaluations are collected for self-improvement purposes, biases can be addressed by collecting additional data or dismissing the results. When used for personnel decisions it is important that any possible bias to student evaluations is empirically studied and controlled for (Braskamp, Brandenburg, & Ory, 1984; Centra, 1993).

Predictive Student Evaluation Variables

Administration of Evaluations

Timing of evaluation. The consensus amongst the relevant literature is that the timing of course evaluation, the date of data collection within the course calendar, does not significantly affect a student’s evaluation of instructor effectiveness (Cashin, 1988; Feldman, 1979; Wachtel, 1998). Costin (1968) compared the mean evaluation scores of graduate teaching assistants collected at the middle of the semester to ratings collected at

the end of the semester and found no statistically significant difference. Frey (1976a) found that evaluation results collected during the last week of an introductory Calculus course were not significantly different than those collected during the first week of the following term. Canaday, Mendelson, and Hardin (1978) employed a similar method, administering course evaluations to random groups of students at one of three time points: (a) preceding the final exam, (b) immediately following the final exam, and (c) after receiving course grades. Results indicated that there were no significant differences in evaluation scores between the groups. Similarly, Marsh and Overall (1980) administered course evaluations at the mid-term and end of term and found that the evaluation scores were highly correlated.

In contrast to the previous studies, Witt and Burdalski (2003) found that evaluations administered on the last day of the semester were lower than at the 11-week mark of a 14-week semester despite self-report data from the student sample that stated opinions of instructor effectiveness were the same or higher than when tested previously. Other contradictory findings come from Aleamoni (1981) and Braskamp et al. (1984), who suggest that evaluation results may be affected if administered before or after an examination. Braskamp et al (1984) report that student ratings collected during the final examination are lower than ratings collected during the semester and recommend administering the student evaluation instrument during the final two weeks of the semester.

The contradictory results are a common thread in the student evaluation literature. Much of the literature pertaining to the potential biasing variables reviewed in the

following sections provides inconclusive evidence of the presence or absence of such a bias. This simply reinforces and justifies the need for additional research in the area.

Anonymity of student raters. Research has been conducted that has found that students that identify themselves (non-anonymity) tend to provide more favorable evaluations than those students that remain anonymous (Blunt, 1991; Braskamp & Ory, 1994; Centra, 1993; Feldman, 1979). Feldman (1979) conducted a review of the published literature that compared student ratings of instructors by students that had identified themselves and students that had completed the evaluation anonymously. In each study, either the same students were used in both the anonymous and nonanonymous conditions or data were gathered from two equivalent sets of students evaluating the same instructor. Of the 10 studies reviewed, seven reported that the nonanonymous student ratings were higher than the anonymous ratings, while three reported little or no difference between the two conditions. Feldman explains that the context of the studies that reported no difference should be considered. The ratings in several of the studies were conducted in an experimental session in which there was no real instructor, thus divorcing the study from the true student and instructor condition.

The anonymity of student raters is largely assured by evaluation administration procedures. It would be considered unethical for an instructor to be able to identify a student's responses. Even with non-identified evaluation forms, students may still doubt the assurance of anonymity (Wachtel, 1998). These concerns may be alleviated by removing the instructor from the room during the administration of the evaluation and having any handwritten responses transcribed by a third-party.

Instructor presence in classroom. Similar to non-anonymity, instructor presence in the classroom tends to lead to more favorable student evaluations of instructor effectiveness (Braskamp & Ory, 1994; Centra, 1993; Feldman, 1979). As part of his review of the literature on circumstances under which evaluations were administered, Feldman (1979) cited a handful of studies that explored the effect of an instructor's presence in the classroom on student ratings. In both studies that Feldman cited student ratings of instructor effectiveness were higher when administered in the presence of the course instructor than in the presence of a neutral party. Page (1974) administered student evaluations to 10 undergraduate psychology courses under two conditions, in the presence of the course instructor and in the presence of a neutral observer. Results indicate that student ratings were higher when administered in the presence of the course instructor. Cooke (1952) found that student ratings were higher when the evaluation instrument was administered in the presence of the instructor than when administered by a student in the class. The difference in scores, however, was not statistically significant. Most researchers suggest that evaluations should be distributed, collected, and scored by a third-party with the evaluated instructor completely removed from the entire evaluation process (Braskamp & Ory, 1994; Centra, 1993).

Course Characteristics

Electivity. Prior research has shown that teachers of elective courses tend to receive slightly higher evaluation scores than teachers of required courses (Centra & Creech, 1976; Feldman, 1978; Gage, 1961; Lovell & Maner, 1955). As part of his review of the published literature on the effect of course characteristics on student ratings of instructor effectiveness, Feldman (1978) cited seven articles that examined the

differences in student ratings for teachers of required and elective courses. Five of the seven articles reviewed (71%) reported that teachers of elective courses received higher ratings than teachers of required courses. The two dissenting articles reported no significant relationship. Centra (1993) suggests that this rating disparity may be due to subdued instructor and student interest in the required courses, which are often introductory in nature and mandated by the university.

Instead of considering electivity as an instructor-level variable as seen above, some researchers have posited electivity as a class-level variable, calculating the percentage of enrolled students in a class taking the course as an elective. Pohlman (1975) did just that in his examination of the effects of several class-level variables, including electivity, on student ratings of instructor effectiveness in a sample of 1,247 university courses with responses from over 33,000 students. The resultant positive zero-order correlations (range of r from .12 to .27) indicate that classes with a high percentage of enrolled students taking the course as an elective tend to have higher ratings than those classes with lower proportions of elective enrollments. Feldman (1978) cited 10 articles that supported this positive relationship, with the resultant relationships of small to moderate strength.

Another way of looking at this phenomenon is by measuring the average intrinsic interest of students in the course. Many student evaluation instruments contain an item that asks the student to rate their interest in the course content at the beginning of the course. This type of item in some studies is used as an indicator of students' intrinsic interest. In Feldman's (1978) systematic review of the literature on the topic, he cited five studies that explored the relationship between intrinsic student interest and student

ratings. The consensus finding was a small positive relationship (r 's in the .10s to .20s) between class ratings and intrinsic student interest.

Class meeting time/length. There currently exists limited research on the relationship between class meeting time (i.e. morning, afternoon, evening) and student ratings of instructor effectiveness. In his synthesis of the existing literature on several course characteristic variables, Feldman (1978) considered the role of course meeting time on student ratings. Of the 11 studies cited, seven concluded that there is no relationship between class meeting time and student rating. Four studies reported slight differences in student ratings amongst the class meeting times but without a consistent pattern across those studies.

Yongkittikul, Gilmore, and Brandenburg (1974) compared student ratings of instructor effectiveness across nine separate course times spanning from 8:00 AM through 4:00 PM. ANOVA results for the overall evaluation score were statistically significant, $f = 1.99$, $p < .05$, with a weak effect size ($\omega^2 = .01$) and nonsignificant post-hoc comparisons (Scheffe). Though the results of Yongkittikul et al. were statistically significant, there was little practical significance. These findings provide further evidence of a nonsignificant relationship between class meeting time and student ratings of instructor effectiveness.

Similar to class meeting time, there is limited research on the relationship between class length and student ratings of instructor effectiveness. As part of a larger study on the optimal class length for undergraduate marketing courses, Reardon, Payan, Miller, and Alexander (2008) examined the relationship between class length and student evaluations of instructor effectiveness in a sample of 1,179 business courses taking place

over a period of 5 years. ANOVA results indicated that class format had a statistically significant effect on student evaluations ($F = 7.40, p = .001$). Instructors teaching short (1 hour/ 3 times a week) and intensive (3 hours/ 1 time a week) format classes received lower evaluation ratings than those teaching moderate (1 ½ hours/ 2 times a week) format classes. Despite the significance of the results, it is difficult to generalize these findings because of the lack of replication and scarce literature on this topic.

Class size. There is empirical evidence that suggests that smaller classes tend to receive higher ratings than larger ones (Feldman, 1984; Neumann, 2000; Davies, Hirschberg, Lye, Johnston, & McDonald, 2007). Feldman (1984) conducted a meta-analysis of the existing literature examining the relationship between class size and student evaluation ratings. Of the 52 articles included in the sample, two reported a positive relationship, 22 reported a statistically significant inverse relationship, 22 reported no relationship, and 11 reported a curvilinear relationship.

This curvilinear relationship is often characterized as U-shaped with small and large classes receiving more favorable ratings than medium-sized classes (Gage, 1961). Centra and Creech (1976) provide evidence of this U-shaped curve in their examination of the mean instructor ratings for nearly 5,000 university classes. The authors found that classes with 15 or fewer students had the highest ratings. Classes with 16 to 35 students and classes with more than 100 students exhibited equivalent ratings; lower than the ratings for the smallest classes. The medium-sized classes (35 to 100 students) ranked the lowest. Pohlman (1975) provided further evidence of the curvilinear relationship between class size and student ratings, reporting statistically significant ($p < .01$) relationships

between class size and the general course, orientation, and presentation of material evaluation subscales.

Centra (1993) proposed several explanations for the higher ratings in the largest classes than the medium-sized classes. The largest classes may be assigned to instructors with the greatest ability to teach large groups of students. Those instructors teaching large classes may spend additional time rehearsing and preparing to teach a large group. And finally, at many universities the large lecture format classes include group seminars, with additional instruction from the professor or teaching assistant.

Workload/rigor. There exist conflicting reports on the effect of course workload on student evaluation scores. The majority of published studies report evidence of a positive relationship between the course workload/rigor and student ratings of instructor effectiveness, with students rating more rigorous courses higher than those with light workloads (Cashin & Slawson, 1977; Marsh & Overall, 1979; Marsh, 1982; Marsh, 1987). Marsh and Overall (1979) found statistically significant positive correlations between student ratings of course workload/difficulty and overall course ($r = .26$) and instructor effectiveness ($r = .16$) in a sample of 186 undergraduate courses. Cashin and Slawson (1977) reported positive correlations, ranging from .14 to .29, between course workload items and those items evaluating instructor effectiveness on the 37 item IDEA evaluation instrument.

In contrast to these findings are results from a few studies that suggest that workload is negatively correlated with student evaluation ratings, meaning that courses with a lighter workload receive higher evaluation scores than more rigorous courses (Greenwald & Gillmore, 1997; Pohlman, 1975). It should be noted that in these studies

correlations were either not reported or extremely small ($r = .02$; Pohlman, 1975). This conceptualization of the relationship between workload and student evaluation is reflected in the idea that expected grade affects student evaluation scores, with a lighter workload often perceived by students as an “easier” grade, and thus the course would be rated more favorably.

Instructor characteristics

As a prelude to the discussion of the potential bias of instructor characteristics, several survey studies that have examined the opinion of faculty members related to the use of student evaluations and potential biases are reviewed. Ryan, Anderson, and Birchler (1980) surveyed 300 faculty, of which 193 (63%) responded, to gather information about their perceptions of the university’s use of student evaluations. Thirty-one percent of respondents reported that they felt undergraduate students were incapable of evaluating instructor performance, 50% felt undergraduates were somewhat capable, with only 17% feeling that undergraduates were “quite capable” or “very capable” of evaluating faculty performance. When asked how the collection of evaluation information has affected general faculty morale, 93.7% of respondents reported morale was somewhat or greatly decreased. About 73% of respondents reported their own morale had somewhat or greatly decreased. Nearly 75% of surveyed faculty reported that the collection of evaluation information had somewhat or greatly increased the distance between faculty and administration. A similar number of faculty (71.5%) reported that their evaluation of university administration had somewhat or greatly decreased in the past several years as a result of the use of student evaluations.

Faculty were asked to assess how the information obtained from student evaluations changed or modified their instructional activities. Thirty-seven percent of respondents reported that the difficulty level of the course had somewhat or greatly decreased as a result of student input. Similarly, 32.6% reported the difficulty of examinations had somewhat or greatly decreased. Twenty-two percent of surveyed faculty reported that amount of material covered in the course had somewhat or greatly decreased because of student input (Ryan, Anderson, & Birchler, 1980).

When asked about the use of student evaluation information, nearly 80% reported that they agreed or strongly agreed with the use of results for the improvement of instruction. Forty-five percent of faculty disagreed or strongly disagreed with the use of student evaluation information for retention and tenure decisions, 44% disagreed or strongly disagreed with the use of student evaluation information for promotion decisions, and 57% disagreed or strongly disagreed with the use of student evaluation information for merit pay decisions (Ryan, Anderson, & Birchler, 1980)..

Birnbaum (2000) surveyed 208 faculty members (68 were untenured) at California State University, Fullerton, to assess their beliefs/opinions about student evaluations. When asked how raising course standards would affect evaluations, 65.4% indicated that higher standards would result in lower evaluation scores. Similarly, 65.9% reported that increasing the amount of course content would result in lower student evaluation ratings. The majority of surveyed faculty reported feeling that the current incentive and promotion structure leads faculty to lower course standards and water down courses. Asked if the current system of promotion and tenure gives incentives to raise

standards of grading, 92.3% responded no. When asked if the use of student evaluations encourages faculty to “water down” course content, 72.1% said yes.

Marsh (1987) surveyed faculty at a major research university and found that faculty members, in the following percentages, believed the factors would bias student ratings: (a) course difficulty, 72%, (b) grading leniency, 68%, (c) instructor popularity, 63%, (d) student interest in subject before course, 62%, (e) course work load, 60%, (f) class size, 60%, (g) student reason for taking course, 55%, and (h) student GPA, 53%.

The above findings suggest that the majority of faculty members are hesitant to completely endorse student course evaluations because of potential biases such as the influence of course workload, course difficulty, and grading leniency. There is evidence of opposition to the use of student evaluation data as the sole indicator for personnel decisions. Most faculty are, however, in favor of using student evaluation data for course improvement, with many faculty reporting changes in their instructional structure based on student feedback.

Instructor experience. Feldman (1983) conducted a meta-analysis of the literature pertaining to the relationship between several variables that measure instructor experience and student evaluation ratings. Thirty-three studies were identified that specifically looked at the relationship between academic rank, usually categorized as instructor, assistant professor, associate professor, and full professor, and overall evaluation. The majority of those studies ($n=22$) concluded that there was no relationship between rank and student rating. Ten studies reported a positive and statistically significant relationship, meaning that instructors with higher rank were rated higher on student evaluations than those instructors of lower rank. Feldman noted that the

magnitude of the positive relationships were relatively weak, with reported correlation coefficients ranging from .06 to .26. One can conclude that there is evidence of a minimal relationship between instructor rank and student evaluation.

Several other studies found nonsignificant differences in student evaluations amongst instructors, assistant professors, associate professors, and full professors but found that ranked college instructors were rated higher than graduate teaching assistants (Aleamoni, 1976; Brandeburg, Slinde, & Batista, 1977; Centra & Creech, 1976). Additionally, research has shown that first-year instructors receive lower evaluation scores than teachers with more experience (Centra, 1978). These findings are more consistent but less worrisome than differences amongst ranked instructors, as teaching assistants and first-year instructors are still learning how to teach and are expected to improve in subsequent years (Centra, 1993).

Two other variables used to assess instructor experience are age of the instructor and the number of years teaching experience, often quantified as the number of years since receiving a doctorate degree. With regard to the relationship between instructor age and student's evaluation ratings, 12 articles were identified; of which six found no relationship and six found a statistically significant inverse relationship. Similarly, surprising results were found when Feldman (1983) analyzed 16 articles that examined the relationship between instructional experience and student ratings. The majority of articles ($n = 9$) found no statistically significant relationship, with two reporting a statistically significant positive relationship and five reporting a statistically significant inverse relationship. While most of these articles found no relationship, it is interesting to note the presence of the inverse relationships between the variables. Those articles found

that older instructors and instructors with more instructional experience were rated lower on student evaluations than their younger and less experienced colleagues.

While there are some conflicting findings related to instructor experience and student ratings of instructor effectiveness, the majority of articles reviewed reported a nonsignificant relationship between the variables.

Gender. There is varied opinion on the relationship between gender and evaluation (Andersen & Miller, 1997; Centra, 1993; Feldman 1992, 1993; Wachtel, 1998). Feldman (1992, 1993) published a two-part meta-analysis examining the effect of gender on student ratings of instructor effectiveness; part one reviewed 485 laboratory and experimental studies while part two focused on studies conducted in real-life settings. Of the laboratory/experimental studies reviewed by Feldman (1992), the vast majority reported that there was not a statistically significant relationship between the instructor's gender and the student's overall evaluation of instructor effectiveness. Those studies that did report a significant effect of gender found that male instructors were rated more favorably than female instructors (Feldman, 1992).

In part two of his meta-analysis, Feldman (1993) reviewed 39 studies conducted in natural settings. Twenty-eight (72%) of the studies reported a product-moment correlation coefficient as the measure of the relationship between instructor gender (coded 0 for male and 1 for female) and student rating, with a positive correlation signifying that females were rated higher than males. Of the 28 studies reporting a correlation coefficient, the majority were positive ($n = 17$), of which eight were statistically significant at the .05 alpha level. Ten of the studies reported negative correlations, signifying that males were rated higher than females, of which 3 were

statistically significant. The mean of the reported correlations ($n = 39$) was .02, which was statistically significant at the .05 alpha level. Though the mean correlation was statistically significant, it is of minimal practical significance, explaining 4/10 of 1 percent of the variance in overall ratings of student evaluations. The combined results of Feldman's two meta-analysis studies reveal that while there are examples of studies that report significant effects of gender on student ratings, the majority of studies found a nonsignificant effect or a statistically significant effect with minimal practical significance (small effect size).

Ethnicity. In his review of the literature on student evaluations, Watchel (1998) cited instructor race/ethnicity as a variable that had not been thoroughly investigated as it relates to student ratings of instructor effectiveness. There is currently minimal research on this variable and its relationship to student evaluation. Ludwig and Meacham (1997) explored the relationship between instructor race and student course evaluations with a sample of 190 undergraduate students. The findings revealed that there was not a statistically significant relationship between instructor race and student ratings.

Shapiro (1990) utilized a similar method with a sample of 399 classes. The zero-order correlation between instructor race (coded 1 for white and 0 for nonwhite) and overall student evaluation was statistically significant, $r = .12$, $p < .001$. The positive correlation indicates that students rated white instructors higher than nonwhite instructors. While the correlation was statistically significant, it is of minimal practical significance, with instructor race explaining about 1% of the variance in overall ratings of student evaluations.

In a more recent study, Bavishi, Madera, and Hebl (2010) examined the effects of instructor ethnicity on student evaluations by presenting 375 high school students (9th - 12th grade) with the following experimental scenario: students imagined themselves in a scenario in which they had received an acceptance letter with a full scholarship to their top choice university. As part of the scholarship they were to work with a professor as a research assistant. The students were asked to review the professor's curriculum vitae and respond to three scales to assess their perceptions of the instructor's competence, interpersonal skills, and legitimacy. Ethnicity was manipulated by the use of name and membership in race specific organizations. The three ethnicities under investigation were Caucasian, Asian-American, and African-American. Results indicated that African-American professors were perceived more negatively on competence, interpersonal skills, and legitimacy than Asian-American and Caucasian professors. Asian-American professors were evaluated comparable to Caucasian professors in the competence and legitimacy dimensions but lower on interpersonal skills. While the study did not directly make use of student course evaluations, the instruments used in the study did measure many of the same aspects or dimensions of instructor effectiveness.

Because of the limited published research on the relationship between instructor ethnicity and student ratings of instructor effectiveness, it is difficult to make any conclusion about the true nature of this relationship. Further research needs to be conducted in this area.

Student Characteristics

Prior interest in subject. As stated previously in the course electivity section, the consensus amongst the research is that a student's prior interest in the course topic is

positively related to his or her ratings of instructor effectiveness (Feldman, 1978; Marsh & Cooper, 1981; Prave & Baril, 1993; Wachtel, 1998). Greater student interest in the subject is related to higher evaluation scores. Prave and Baril (1993) suggest that it may be necessary to control for this bias, particularly for general education subjects like math and English. As a result many student evaluation instruments, including the one investigated in this study, now contain direct measures of a student's prior interest in the course topic.

Gender. Similar to instructor gender, there are conflicting findings related to student gender and evaluation (Andersen & Miller, 1997; Centra, 1993; Feldman 1992, 1993; Wachtel, 1998). As part of Feldman's (1992, 1993) two-part meta-analysis examining the effect of gender on student ratings of instructor effectiveness, the interaction effect of instructor and student gender were examined. The vast majority ($n = 28$) of experimental/laboratory studies that examined the interaction effect ($n = 31$) of student and instructor gender on student evaluations found no significant interaction. The three published studies that reported a statistically significant interaction effect found a same-gender bias with male students rating male instructors higher than female instructors and female student rating female instructors higher than male instructors (Feldman, 1992).

In his review of classroom studies, Feldman (1993) identified ten studies that examined the interaction effect of student and instructor gender on student evaluation ratings. Each study rank ordered four categories from highest (1) to lowest (4) evaluation ratings: (a) female rating female, (b) male rating male, (c) female rating male, and (d) male rating female. Based on the results from the first meta-analysis, it would be

expected that the average ranking for same-gender ratings (categories a and b) would be higher than average ranking for cross-gender ratings (categories c and d). However, in contrast to the findings from the classroom studies, results show that the average rank order for same-gender categories were 2.00 and 2.05 and the average rank order for cross-gender ratings were 2.95 and 3.00.

In a more recent study, Centra and Gaubatz (2000) examined the interaction between student and instructor gender on student ratings of instructor effectiveness in a sample of 741 classes from 21 postsecondary institutions. MANOVA results indicated the presence of a gender bias for female instructors. Female students rated female professors higher than male students on five of the seven evaluation scales. There was no significant difference in how male instructors were rated. The conflicting findings about the interaction effect of student and instructor gender on evaluation make it difficult to make any conclusive statement about the nature of the relationship.

Student age. In his review of the existing literature on student course evaluations, Wachtel (1998) noted that there had not been any recent studies that explored the relationship between student age and course ratings, stating that further studies needed to be conducted. The existing literature on the relationship between student age and course rating provides conflicting evidence on the directionality of the relationship. A study conducted to assess the relationship between student grade and course evaluation rating included student age as a control variable. Age was removed from the final equation because it was found not to vary systematically with student course rating (Seiver, 1983). Another study utilized age as a control variable in an analysis of the relationship between an instructor's extraversion and course rating. The zero-order correlation between student

age and student rating of instructor effectiveness was statistically significant at the .05 alpha level ($r = .17$; Radmacher & Martin, 2001). These results suggest a weak positive relationship between student age and course rating, with older students providing higher ratings of instructor effectiveness than younger students. In contrast to the studies reporting a positive relationship between age and student ratings are the results of Worthington (2002) who reported a negative relationship between age and student course rating, with students over 30 years of age providing lower evaluation ratings than students between the ages of 21 and 30. Given the scarcity of published research and conflicting nature of existing findings, further research should be conducted to assess the nature of this relationship.

Expected grade. There is evidence of a positive correlation between student's expected course grade and evaluation of instructor effectiveness. A common criticism of student evaluations is that the instrument is more a measure of student's satisfaction with expected grade than a true reflection of instructor effectiveness. Review of literature on the subject from 1970 to the present does not provide a clear univocal relationship. Many research studies suggest that there is a positive correlation between expected grade and evaluation score (Braskamp & Ory, 1994; d'Appolonia & Abrami, 2007; Marsh, 1987; Marsh & Dunkin, 1992; McPheson & Jewel, 2007).

In a review of the literature published on this topic between 1924 and 1998, Aleamoni (1999) cited 37 studies that have reported a statistically significant positive relationship between actual or expected course grade and student ratings of instructor effectiveness. Twenty-four studies reported no relationship between grade and student rating, and one study reported a statistically significant negative correlation. The mean

correlation across the 62 studies was 0.18 ($Mdn = 0.14$, $SD = 0.16$) which is indicative of a relatively weak relationship. Other meta-analytic studies have cited a mean correlation between .10 and .30 (Centra, 2003; Feldman, 1997), a relationship of moderate magnitude.

As part of a larger study on the relationship between course, instructor and student characteristics, and student evaluations, Kozub (2010) explored the effect of expected course grade on evaluation ratings in a sample of 463 undergraduate business majors. The correlation between overall evaluation and expected course grade was statistically significant at the .001 alpha level, $r = .36$, meaning that about 13% of the variance in students' overall rating of instructor effectiveness was accounted for by the expected course grade. The correlations between expected course grade and individual dimensions/subscales of the evaluation scale were also statistically significant (pedagogical dimension $r = .15$, rapport dimension $r = .24$, difficulty dimension $r = .35$, and value dimension $r = .26$).

One of the larger studies on the effect of expected course grade on student evaluation ratings effect was conducted by Centra and Creech (1976). Utilizing a sample of 9,194 class-average ratings from several diverse universities, Centra and Creech calculated a correlation of .20 between student rating and expected course grade. Pohlman (1975) also utilized a large sample, 1,247 class-average ratings from Southern Illinois University, Carbondale, in his exploration of the relationship between expected course grade and student ratings. Utilizing the 40 instructor and course evaluation items from the Instructional Improvement Questionnaire (IIQ), Pohlman calculated a correlation of .42 between expected course grade and general course rating.

Despite the weak to moderate magnitude of the relationship, the fact is that there is consistent evidence of a positive relationship between expected course grade and student evaluation score. The point of contention is how to interpret these findings. Marsh (1987; Marsh & Roche, 1997) provides three possible interpretations: (a) the leniency hypothesis, which posits that instructors with more lenient grading standards receive higher student evaluation ratings, (b) the validity hypothesis, which implies that effective instructors cause students to work harder, learn more and earn better grades, and (c) the student characteristic hypothesis, which suggests that pre-existing student characteristics such as prior subject interest influence both teaching effectiveness and student evaluation scores. Marsh and Roche (1997) warn that many of the experimental studies that claim to demonstrate the grading leniency effect are methodologically flawed and biased. In contrast, Haskell (1997) argues that the same studies provide evidence of the grading leniency effect and suggests that there have been efforts undertaken by many evaluation researchers to hide this conclusion. There exists very spirited debate on both sides of this issue and no consensus explanation.

Summary

There exists a substantial body of literature on student evaluations of instructor effectiveness. Current estimates of the amount of published research on student evaluation of instructor effectiveness range from 1300 (Cashin, 1995) to 2000 (Feldman, 2003) citations. The research is so vast that periodically reviews of existing literature are published. Feldman (2003) cites 34 such reviews, having authored 15 meta-analyses of the relevant literature himself. Within this paper, the citations range in publication date from 1928 to 2010, covering a span of over 80 years.

While many researchers contend that scores obtained from current student evaluation instruments are valid and reliable measures (Aleamoni, 1999; Cashin, 1995; Centra, 1993; Costin, Greenough, & Menges, 1971; Firth, 1979 Marsh, 1984; Marsh & Overall, 1980) there is still a large number of researchers that contend that the results from such instruments should not be relied upon for making personnel and tenure decisions (Wachtel, 1998). Opponents cite several points of contention: (a) there is not a consensus definition of effective teaching, (b) teaching to promote positive course evaluations may conflict with good teaching practice, (c) and evaluation scores may be influenced by variables (biases) that have nothing to do with instructor effectiveness (Wachtel, 1998). When used for personnel decisions it is critical that any possible bias to a student course evaluation score is empirically studied and controlled for (Braskamp, Brandenburg, & Ory, 1984; Centra, 1993).

The majority of the published literature on student ratings of instructor effectiveness is focused on the exploration of the potential student, course and instructor-level biases. Table X provides a summary of the relationships between student ratings of instructor effectiveness and the potential biasing variables reviewed in this chapter. Despite the abundance of empirical research on the relationship between these potential biasing characteristics and student ratings, there remains a great deal of uncertainty about the true nature of these relationships. Contradictory results are a common thread in much of the student evaluation literature resulting in inconclusive evidence of the presence or absence of such a bias.

Table X

Summary of the variables influencing student ratings

Potential biasing variable	Summary of the relationship/effect
Administration of evaluation	
Timing of evaluation	Inconclusive: contradictory findings
Anonymity of rater	Students that identify themselves (non-anonymity) tend to provide more favorable evaluations than those students that remain anonymous (Blunt, 1991; Braskamp & Ory, 1994; Centra, 1993; Feldman, 1979).
Instructor presence in classroom	Instructor presence in the classroom at the time of the evaluation tends to lead to more favorable student evaluations of instructor effectiveness (Braskamp & Ory, 1994; Centra, 1993; Feldman, 1979).
Course characteristics	
Electivity	Teachers of elective courses tend to receive slightly higher evaluation scores than teachers of required courses (Centra & Creech, 1976; Feldman, 1978; Gage, 1961; Lovell & Maner, 1955).
Class meeting time/length	Inconclusive: limited published literature
Class size	Empirical evidence suggests that smaller classes tend to receive higher ratings than larger ones (Davies et al., 2007; Feldman, 1984; Neumann, 2000). Potential curvilinear relationship (Centra & Creech, 1976; Gage, 1961; Pohlman, 1975)
Workload/rigor	Inconclusive: contradictory findings
Instructor characteristics	
Instructor experience	Nonsignificant relationship between experience and student ratings (Feldman, 1983)
Instructor gender	Inconclusive. While there are examples of studies that report significant effects of gender on student ratings, the majority of studies found a nonsignificant effect or a statistically significant effect with minimal practical significance (Feldman 1992, 1993).
Instructor ethnicity	Inconclusive: limited published literature
Student characteristics	
Prior interest in subject	Students' prior interest in the course topic is positively related to their ratings of instructor effectiveness. (Feldman, 1978; Marsh & Cooper, 1981; Prave & Baril, 1993; Wachtel, 1998).
Gender	Inconclusive. Potential same gender bias. Males rate males higher. Females rate females higher (Feldman, 1992, 1993)
Expected grade	Positive correlation between expected grade and evaluation score (d'Appolonia & Abrami, 2007; Marsh, 1987; Braskamp & Ory, 1994, Marsh & Dunkin, 1992, McPheson & Jewel, 2007).

The major purpose of this study (RQ1) and framework underlying the presentation of this paper is the construction of an empirically based prediction model (Figure 1) of student ratings on instructor effectiveness with the student, class, and instructor-level variables most prevalent in the body of student evaluation literature. This model aims to both replicate the findings for those variables with well-defined relationships and explicate those relationships with inconclusive findings.

Research Questions

The research questions being addressed in this study are as follows:

1. Do the scores obtained in the study exhibit adequate reliability and construct validity?
2. What student, course and instructor-level variables are statistically significant predictors of student ratings of instructor effectiveness?

Student-level independent variables: gender, ethnicity, age, prior interest in subject, and course electivity.

Class-level independent variables: size (enrollment), meeting time, length, and level (undergraduate/graduate).

Instructor-level independent variables: gender, ethnicity, rank/position.

Dependent variable: student rating of instructor effectiveness.

3. Do students' ratings of instructor effectiveness significantly predict students' final course grades?

Figure 1

Prediction model

CHAPTER III

METHODS

The topics presented in this chapter include a statement of the research questions, a description of the study participants, the instruments and their psychometric properties, the procedures, and the data analysis methods.

Participants

The study took place in the context of the College of Education and Human Development at a large metropolitan research university in the southern United States. The course evaluation data in this study were naturally nested in the format of students within classes within instructors. Relevant descriptive statistics that illustrate participant characteristics, such as frequency, mean, standard deviation and range, are provided at the student, course, and instructor-levels.

All of the course evaluation data from the College of Education and Human Development for that were taught during the fall 2010 and spring 2011 semesters were included in the sample with a few exceptions: (a) courses that were taught by multiple instructors, (b) courses with an enrollment of one student, and (c) online courses. Course with multiple instructors were removed because it was unclear how course delivery was divided up amongst the instructors. Without knowing how course instruction was divided amongst the instructors it was difficult to accept the validity of those scores and for that reason those courses were excluded. Courses with a singular student enrollment were also removed due to the potential biases due to a student's lack of anonymity. It is worth

noting that results of course evaluations with five or fewer completed surveys were withheld from the instructor due to the threats to confidentiality. Online courses were not included in the sample because at that time the data were collected, the online courses in the College of Education and Human Development utilized a different course evaluation instrument than courses with face-to-face content delivery.

The remaining courses yielded an initial sample of 5,629 course evaluations. Five hundred and sixty-nine course evaluations were missing on the entirety of student demographic and grade variables and were removed from the sample leaving a sample of 5,060 course evaluations. The decision was made to impute predicted values for the cases with missing data on the student level predictor variables. Prior to the use of multiple imputation several additional cuts were made to the data. There were 249 student evaluations that had incomplete grades or were graded on a pass/fail scale. Because of the interest in the relationship between final course grade and student evaluation of perceived instructor effectiveness those courses which did not grade on a 4.0 scale (i.e. pass/fail) were removed. Eighty-nine evaluations were missing on 20% ($n = 5$) or more of the course evaluations items and were not included in the imputation.

Participant demographics

These reductions left a sample of 4722 student course evaluations nested within 572 classes nested within 203 instructors that were missing on less than 20% of course evaluation items. A final data reduction was made to remove those cases with missing course or instructor predictor data ($n = 490$). The decision was made not to impute values for cases missing the course level variables because of the limited number of covariates that could be used to estimate plausible values. All of the missingness at the course level

occurred in the schedule related variables (start time, length, meeting pattern), such that any case with missing data at the course level was missing values for all three schedule related variables. If employed, imputation methods would predict start time, length, and meeting pattern based on values from enrollment and level (undergraduate/graduate). Similarly, at the instructor level, missingness occurred within the gender and race variables, such that any case with missing data at the instructor level was missing on gender and race, leaving instructor title to estimate missing values. It is because of the limited number of covariates by which to estimate plausible values that the decision was made to use listwise deletion for missingness at the course and instructor levels.

Table 3

Sample demographics: Means and standard deviations

Variable	Mean	Standard Deviation
Age	28.36	9.49
Prior interest	3.61 (2.61 ¹)	1.00
Effort	4.07 (3.07 ¹)	0.82
Course enrollment	19.10	11.61
Course length (minutes)	136.77	44.62

A reduced sample of 4232 student responses nested within 475 classes within 178 instructors was used to answer research questions two and three. As shown in Table 4, the majority of students were female (73.1%) and white (84.1%). The average student age was around 28 years (SD = 9.29).

¹ This value was calculated with the variable centered at 0 for HLM analysis, coded from 0 to 4.

The majority of courses (n = 475) began in the late afternoon, between 4:01 and 6:00 PM (52.6%). The remaining 225 courses were spread rather evenly throughout the day: (a) 8: 01 to 10 AM (18.1%), (b) 10:01 to 12 PM (5.9%), (c) 12:01 to 2PM (13.5%), (d) 2:01 to 4PM (4.4%), and (e) 6:01 to 8PM (5.5%). Average course enrollment was 19.10 and an average length of 136.7 minutes (about two and a half hours). Close to 80% of the classes met once a week. The classes were almost evenly split among the undergraduate (45.3%) and graduate (54.7%) levels.

Table 4

Sample demographics: Frequencies for binary variables

Variable	Frequency of 0 (%)	Frequency of 1 (%)
Student gender	1138 (26.9%) Male	3094 (73.1%) Female
Student race	3557 (84.1%) White	675 (15.9%) Non-white
Electivity	298 (7 %) Not required	3934 (93%) Required
Class level	215 (45.3%) Undergrad.	260 (54.7%) Grad.
Class meeting pattern	373 (78.5%) 1 day	102 (21.5%) > 1 day
Instructor gender	65 (36.5%) Male	113 (63.5%) Female
Instructor race	155 (87.1%) White	23 (12.9%) Non-white

The 178 instructors exhibited similar demographics to the student respondents, with the majority of instructors reporting that they are white (87.1%) and female (63.5%). Most of the course instructors were listed as instructor, lecturer, or staff (57.9%); nearly 19% were listed as an associate professor, almost 12% were listed as a full professor, 8% as assistant professor, and 4% were listed as a graduate assistant.

Instruments

Student Course Evaluation

The university at which the current study was conducted made a university-wide transition to online student evaluations in the fall of 2010. Students received an email from the dean of the college during the last month of the semester asking them to complete the course evaluations for the courses in which they were enrolled. The email stated that student feedback would help the college to assess the instructional quality of the coursework and identify opportunities for improvement and that all responses would remain anonymous. Students were able to complete the course evaluation at any time from the receipt of the email until the end of the semester. Prior to the conversion to the online format the evaluation instrument had been distributed during a class meeting in the paper-and-pencil format.

Student perceptions of instructor effectiveness were assessed with a 25 item evaluation form developed in 1988 by a faculty committee, with input from faculty, students and administrators at the College of Education and Human Development (Petrosko, 1990). The first six items are related to student background, with questions asking about student effort, interest and the electivity of the course. The evaluation component is composed of 19 statements (items 7-25) about the instructor's teaching ability, preparation, grading, the course text and organization to which the student rates their agreement with the statement on a 5 point Likert-type scale ranging from 1 "Strongly Disagree", "Poor", or "Very Low" to 5 "Strongly Agree", "Excellent" or "Very High". A copy of the instrument is located in *Appendix 1*.

A pilot test conducted in the summer of 1988, during the development of the student evaluation instrument, reported Cronbach's alpha of .94 for the 19 evaluation items in a sample of 1334 students. Additional analyses were conducted at the classroom level, with classroom means ($n = 76$ classes) used in place of student scores, revealing a reliability coefficient of .96 (Petrosko, 1990).

Data Source

Additional student, course, and instructor variables were accessed through the university's Office of Academic Planning and Accountability, in accordance with the approved Institutional Review Board protocol. The preexisting data were structured in such a way that a student's course evaluation responses could be linked to student variables such as ethnicity, age, gender, and final course grade. Similarly, course evaluation responses were linked to course level variables such as class size, the time of the course meeting, length of the class meeting, and the level of the course (undergraduate vs. graduate) as well as instructor-level variables such as ethnicity, gender, and the instructor's position.

Variables. In this section, the student, class, and instructor-level predictor variables are defined. As mentioned previously, the variables examined in this study were chosen after a thorough review of the literature in an attempt to build a model for the prediction of student rating of instructor effectiveness.

The student-level variables electivity and prior subject interest were gathered from student responses to the course evaluation instrument (*Appendix 1*). Course electivity was measured by an item on the course evaluation instrument (item number 1) that asks the respondent if the course is required for their program. Response options are

“yes”, “no”, or “not applicable/cannot answer” The variable was treated as a dichotomous variable with “yes” responses coded as 0, “no” responses coded as 1, and “not applicable/cannot answer” responses treated as missing values. Prior subject interest was measured with a retrospective Likert-type item on the course evaluation instrument (item number 2) that asks the respondent to indicate their interest in the course content at the beginning of the course. Response options range on a five point Likert-type scale from “very low” to “very high”.

The remaining student-level variables, ethnicity, age, gender, and final course grade, were gathered from a database maintained by the University. Student ethnicity is a university maintained variable that serves as an indicator of the student’s race. The variable was treated as a dichotomous variable with “white” coded as 0, and “non-white” coded as 1. Age is a university maintained continuous variable that serves as an indicator of the student’s biological age. Gender is a university maintained continuous variable that serves as an indicator of the student’s biological sex with “male” coded as 0, and “female” coded as 1. Final course grade is the grade awarded to the student at the conclusion of the course; an ordinal variable on a 0-4.0 scale.

The class-level variables, class size, the time of the course meeting, length of the class meeting, and the level of the course, were gathered from a database maintained by the University. Class size is a continuous measure of the total enrollment for the class. The time of the course meeting is a university maintained variable that provides the starting time for the course as listed in the university schedule. The variable was treated as an interval variable coded as 0 for 8:00AM to 10:00AM, 1 for 10:01AM to 12:00PM, 2 for 12:01PM to 2:00 PM, 3 for 2:01PM to 4:00 PM, 4 for 4:01PM to 6PM, and 5 for

6:01PM to 8PM. Length of the class meeting is a university maintained variable that provides the length of an individual class session. It was converted from class length in hours to class length in minutes. The level of the course distinguishes between undergraduate (coded 0) and graduate (coded 1) level courses.

The instructor-level variables, ethnicity, gender, and the instructor's rank, were gathered from a database maintained by the University. Instructor ethnicity is a university maintained variable that serves as an indicator of the instructor's race. The variable was treated as a dichotomous variable with "white" coded as 0, and "non-white" coded as 1. Instructor gender is a university maintained continuous variable that serves as an indicator of the instructor's biological sex with "male" coded as 0, and "female" coded as 1. Instructor position is a university maintained ordinal variable that indicates the instructor's position at the university. The variable was coded 0 for graduate assistant, 1 for instructor/lecturer/staff, 2 for associate professor, 3 for assistant professor, and 4 for full professor.

Student course rating is a composite variable made up of 19 course evaluation items about the instructor's teaching ability, preparation, grading, the course text and organization to which the student rated their agreement with the statement on a 5 point Likert-type scale ranging from 1 "Strongly Disagree", "Poor", or "Very Low" to 5 "Strongly Agree", "Excellent" or "Very High". The variable was represented as the sum of student responses to the 19 evaluation items, with possible scores ranging from 1-95.

Procedure

The student course evaluation and additional student, class and instructor-level data were accessed through a secure server at a computer terminal in the Office of

Academic Planning and Accountability in accordance with the approved Institutional Review Board protocol. All unique identifiers of the instructor, student and class were removed from the data prior to the researcher's receipt of the data. At the time of analysis it was impossible for the researcher to identify any student, class or instructor or to link the data in any combination of files that would make identification possible.

Analysis

Research Question One

The first research question is "do the scores obtained in the study exhibit adequate reliability and construct validity?" The reliability of the obtained scores was assessed by calculating appropriate reliability statistics while the construct validity was assessed by conducting a factor analysis.

The construct validity, defined as how well the instrument measures or correlates with the construct under investigation, was assessed by conducting an exploratory factor analysis. Principal axis factoring was used to extract the factors from the data. While both principal axis factoring and principal components analysis (PCA) use the same method for extracting factors from a correlation matrix, principal axis factoring provides an estimate of the communality for each item (Russell, 2002). In comparison, principal components analysis sets the communalities to 1.0, essentially assuming that all of the variance in an item is explainable by the factors that are derived. Principal axis factoring was preferred over PCA because the inclusion of the communality values approximates the analysis of a covariance matrix, where the variance of each item reflects its association with the other items in the factor analysis (Russell, 2002). In addition to theoretical evidence in favor of principal axis factoring, empirical studies have found

principal axis factoring to provide similar or more accurate results than PCA (Bentler & Kano, 1990; Schneewiss, 1997; Widaman, 1993).

To determine the number of factors to retain two methods were used: (a) parallel analysis, and (b) an examination of the scree plot of eigenvalues from the reduced correlation matrix. Parallel analysis is a four step process that involves the comparison of eigenvalues generated from real data with eigenvalues generated from parallel random data. Factors from the real data with eigenvalues greater than the eigenvalues from the random data are retained. Parallel analysis was preferred over the commonly used Kaiser's eigenvalue-greater-than-1 (K1) rule because of the strong empirical evidence in favor of parallel analysis. Several studies have shown that the K1 rule is less accurate than parallel analysis and tends to overestimate the correct number of factors by as much as 66% (Hayton, Allen, & Scarpello, 2004; Linn, 1968; Horn, 1965).

Following the extraction of factors, oblique rotation methods were used to increase the interpretability of the results. Promax rotation with Kaiser normalization was selected because of the expected moderate correlations between the items on the student evaluation instrument and oblique rotations ability to allow items to be correlated with one another. Items with a minimum factor pattern coefficient of .40 were considered appropriate for inclusion in the rotated factors (Stevens, 2001).

The internal consistency of the obtained scores was assessed by calculating Cronbach's alpha. Internal consistency can be defined as the degree to which items on an instrument measure that attribute in a consistent manner (Tashakkori & Teddlie, 1998). It is determined by calculating the average correlation between items on the instrument. Nunally (1978) suggests a minimum reliability of .70. Reliability estimates were

calculated for the each of the emergent factors as well as an estimate of the reliability of the overall student rating instrument.

Research Question Two

The second research question posits, “what are the significant student, course, and instructor-level predictors of student ratings of instructor effectiveness?” The relationship between the predictors and outcome variable were assessed using hierarchical linear modeling (HLM).

In order to test the prediction model, regression analysis was employed in which the predictor variables are regressed on student course ratings. Because this research takes place in a school setting, the regression assumption of independence does not hold true. Students are nested in a multilevel setting of classes nested in instructors, so one cannot expect that the outcome values (student ratings of instructor effectiveness) are independent of their class, instructor, and of other students within their class. In order to account for the non-independence of units, the analysis that was employed was the Hierarchical Linear Modeling (HLM) approach of Bryk and Raudenbush (1992). Analyses were conducted with *HLM v6.08 for Windows*.

There are three stages to the HLM model building process: (a) an unconditional model, (b) a random coefficients model, and (c) a contextual model. Because of the strong empirical basis for the inclusion of the predictor variables, non-statistically significant predictors ($p > .05$) were not trimmed from the model. In addition to theoretical reasons for keeping non-statistically significant predictors there is the fact that when using the iterative process of maximum likelihood estimation, the addition or removal of predictors affects all of the estimates in the model.

Fully unconditional model. An unconditional model was estimated in which only the outcome variable (student rating of instructor effectiveness) was entered into the model without any predictors. In an unconditional model, the variability in overall student course evaluation scores is partitioned into variance between students within classes ($\sigma^2 / \sigma^2 + \tau_\pi + \tau_\beta$), variance between classes within instructors ($\tau_\pi / \sigma^2 + \tau_\pi + \tau_\beta$), and variance between instructors ($\tau_\beta / \sigma^2 + \tau_\pi + \tau_\beta$; Raudenbush & Bryk, 2002). Below are the equations for the fully unconditional model.

$$Y(\text{Eval})_{ijk} = \pi_{0jk} + e_{ijk} \quad (1)$$

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \quad (2)$$

$$\beta_{00k} = \gamma_{000} + u_{00k} \quad (3)$$

$Y(\text{Eval})_{ijk}$ is the course rating of student i in course j for instructor k . π_{0jk} is the mean course rating in course j and instructor k . e_{ijk} is the amount of deviation of student ijk from the course mean. β_{00k} is the mean course rating for instructor k . r_{0jk} is the amount of deviation in course jk from the instructor mean. γ_{000} is the grand mean course rating for the entire sample of students nested in course nested in instructors. u_{00k} is the amount of deviation in instructor k from the grand mean.

Random coefficients model. A random coefficients model was estimated in which overall student course evaluation scores were considered a function of student age, student ethnicity, student gender, course electivity and prior student interest. By comparing the τ estimates between the fully unconditional and random coefficients model, the proportion of variance in student ratings of instructor effectiveness explained by the student-level predictors was calculated. Below are the equations for the random coefficients model. As indicated by the equations, for all of the student-level predictors

other than student gender and race, in which both the intercept and random error vary (9, 10, 16, & 17), only the intercept varied across class and instructor. The interaction between student and instructor gender and student and instructor race were allowed to vary because of empirical evidence suggesting the potential for such a relationship (Bavishi, Madera, & Hebl, 2010; Feldman, 1993, 1993).

$$Y(\text{Eval})_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{electivity}_{jk}) + \pi_{2jk}(\text{interest}_{jk}) + \pi_{3jk}(\text{effort}_{jk}) + \pi_{4jk}(\text{stud_gender}_{jk}) + \pi_{5jk}(\text{stud_race}_{jk}) + \pi_{6jk}(\text{stud_age}_{jk}) + e_{ijk} \quad (4)$$

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \quad (5)$$

$$\pi_{1jk} = \beta_{10k} \quad (6)$$

$$\pi_{2jk} = \beta_{20k} \quad (7)$$

$$\pi_{3jk} = \beta_{30k} \quad (8)$$

$$\pi_{4jk} = \beta_{40k} + r_{4jk} \quad (9)$$

$$\pi_{5jk} = \beta_{50k} + r_{5jk} \quad (10)$$

$$\pi_{6jk} = \beta_{60k} \quad (11)$$

$$\beta_{00k} = \gamma_{000} + u_{00k} \quad (12)$$

$$\beta_{10k} = \gamma_{100} \quad (13)$$

$$\beta_{20k} = \gamma_{200} \quad (14)$$

$$\beta_{30k} = \gamma_{300} \quad (15)$$

$$\beta_{40k} = \gamma_{400} + u_{40k} \quad (16)$$

$$\beta_{50k} = \gamma_{500} + u_{50k} \quad (17)$$

$$\beta_{60k} = \gamma_{600} \quad (18)$$

$Y(\text{Eval})_{ijk}$ is the course rating of student i in course j for instructor k . π_{0jk} is the intercept for course j in instructor k . Electivity is the course electivity predictor variable

(coded 0 for yes and 1 for no). Interest is the prior student interest predictor variable (grand mean centered). StudGen is the student gender predictor variable (coded 0 for male and 1 for female). StudAge is the student age predictor variable (grand mean centered). StudEth is the student ethnicity predictor variable (coded 0 for white and 1 for nonwhite). In the random part of the model, e_{ijk} is the incremental effect of student i in class j for instructor k . r_{0jk} is the incremental effect of class j for instructor k . u_{00k} is the incremental effect of instructor k to the observed outcome.

Contextual model. Class-level. The next stage of analysis was the estimation of a contextual model in which the student-level intercepts were considered a function of the time of the class meeting, length of the class meeting, class size, and the level of the course (undergraduate vs. graduate). By comparing the τ estimates between the random coefficients and class-level contextual model, the proportion of variance in student ratings of instructor effectiveness explained by the class-level predictors above and beyond what was explained by the student-level variables were calculated. Below are the equations for the class-level contextual model. As indicated by the equations, for all of the class-level predictors other than student gender and race, in which both the intercept and random error vary (24, 25, 31, & 32), only the intercept varied across class and instructor.

$$Y(\text{Eval})_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{electivity}_{jk}) + \pi_{2jk}(\text{interest}_{jk}) + \pi_{3jk}(\text{effort}_{jk}) + \pi_{4jk}(\text{stud_gender}_{jk}) + \pi_{5jk}(\text{stud_race}_{jk}) + \pi_{6jk}(\text{stud_age}_{jk}) + e_{ijk} \quad (19)$$

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(\text{enroll}_{jk}) + \beta_{02k}(\text{level}_{jk}) + \beta_{03k}(\text{length}_{jk}) + \beta_{04k}(\text{day}_{jk}) + \beta_{05k}(\text{time}_{jk}) + r_{0jk} \quad (20)$$

$$\pi_{1jk} = \beta_{10k} \quad (21)$$

$$\pi_{2jk} = \beta_{20k} \quad (22)$$

$$\pi_{3jk} = \beta_{30k} \quad (23)$$

$$\pi_{4jk} = \beta_{40k} + r_{4jk} \quad (24)$$

$$\pi_{5jk} = \beta_{50k} + r_{5jk} \quad (25)$$

$$\pi_{6jk} = \beta_{60k} \quad (26)$$

$$\beta_{00k} = \gamma_{000} + u_{00k} \quad (27)$$

$$\beta_{10k} = \gamma_{100} \quad (28)$$

$$\beta_{20k} = \gamma_{200} \quad (29)$$

$$\beta_{30k} = \gamma_{300} \quad (30)$$

$$\beta_{40k} = \gamma_{400} + u_{40k} \quad (31)$$

$$\beta_{50k} = \gamma_{500} + u_{50k} \quad (32)$$

$$\beta_{60k} = \gamma_{600} \quad (33)$$

β_{00k} is the intercept for instructor k in modeling the class effect π_{0jk} . Enroll is the class size predictor variable (grand mean centered). Level is the class-level predictor variable (0 for undergraduate and 1 for graduate level). Length is the class length predictor variable (grand mean centered). Day is the number of days per week class variable (0 for one day and 1 for two or more days). Time is the class meeting time predictor variable.

Instructor-level. The final stage of analysis was the estimation of a contextual model in which class-level intercepts were considered a function of instructor ethnicity, gender, and instructor position. By comparing the τ estimates between the class-level and instructor-level contextual model, the proportion of variance in student ratings of instructor effectiveness explained by the instructor-level predictors above and beyond

what was explained by the class-level variables were calculated. Below are the equations for the instructor-level contextual model. Equation 13 includes instructor gender as a predictor of the student gender slope in an attempt to model the potential interaction effect between instructor and student gender. Equation 14 includes instructor race as a predictor of the student race slope in an attempt to model the potential interaction effect between instructor and student gender. The interaction between student and instructor gender and student and instructor race were allowed to vary because of empirical evidence suggesting the potential for such a relationship (Bavishi, Madera, & Hebl, 2010; Feldman, 1993, 1993). This final model was used for interpreting results.

$$Y(\text{Eval})_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{electivity}_{jk}) + \pi_{2jk}(\text{interest}_{jk}) + \pi_{3jk}(\text{effort}_{jk}) + \pi_{4jk}(\text{stud_gender}_{jk}) + \pi_{5jk}(\text{stud_race}_{jk}) + \pi_{6jk}(\text{stud_age}_{jk}) + e_{ijk} \quad (34)$$

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(\text{enroll}_{jk}) + \beta_{02k}(\text{level}_{jk}) + \beta_{03k}(\text{length}_{jk}) + \beta_{04k}(\text{day}_{jk}) + \beta_{05k}(\text{time}_{jk}) + r_{0jk} \quad (35)$$

$$\pi_{1jk} = \beta_{10k} \quad (36)$$

$$\pi_{2jk} = \beta_{20k} \quad (37)$$

$$\pi_{3jk} = \beta_{30k} \quad (38)$$

$$\pi_{4jk} = \beta_{40k} + r_{4jk} \quad (39)$$

$$\pi_{5jk} = \beta_{50k} + r_{5jk} \quad (40)$$

$$\pi_{6jk} = \beta_{60k} \quad (41)$$

$$\beta_{00k} = \gamma_{000} + \gamma_{001}(\text{InstructRace}_k) + \gamma_{002}(\text{InstructGen}_k) + \gamma_{003}(\text{InstructPos}_k) + u_{00k} \quad (42)$$

$$\beta_{10k} = \gamma_{100} \quad (43)$$

$$\beta_{20k} = \gamma_{200} \quad (44)$$

$$\beta_{30k} = \gamma_{300} \quad (45)$$

$$\beta_{40k} = \gamma_{400} + \gamma_{401}(\text{InstructGen}_k) + u_{40k} \quad (46)$$

$$\beta_{50k} = \gamma_{500} + \gamma_{501}(\text{InstructRace}_k) + u_{50k} \quad (47)$$

$$\beta_{60k} = \gamma_{600} \quad (48)$$

γ_{000} is the intercept term in the instructor-level model for β_{00k} . InstructRace is the instructor ethnicity predictor variable (coded 0 for white and 1 for nonwhite). InstructGen is the instructor gender predictor variable (coded 0 for male and 1 for female). InstructPos is the instructor position predictor variable (grand mean centered).

Research Question Three

The third research question, “do a student’s ratings of instructor effectiveness predict a student’s final course grade?” was also answered by using the Hierarchical Linear Modeling (HLM) approach of Bryk and Raudenbush (1992). The same rationale from question one can be applied to the explanation of the selection of analysis. While the goal of this question is to determine whether a student’s rating significantly predict a student’s final course grade as opposed to considering the collective contribution of several predictors in explaining the variation in student ratings, HLM was chosen because of its ability to control for student, class and instructor-level variables and appropriately parcel out student, class and instructor variation.

Fully unconditional model. An unconditional model was estimated in which only the outcome variable (student course grade) was entered into the model without any predictors. In an unconditional model the variability in student final course grade is partitioned into variance between students within classes ($\sigma^2 / \sigma^2 + \tau_\pi + \tau_\beta$), variance between classes within instructors ($\tau_\pi / \sigma^2 + \tau_\pi + \tau_\beta$), and variance between instructors ($\tau_\beta /$

$\sigma^2 + \tau_\pi + \tau_\beta$; Raudenbush & Bryk, 2002). Below are the equations for the fully unconditional model.

$$Y(\text{Grade})_{ijk} = \pi_{0jk} + e_{ijk} \quad (49)$$

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \quad (50)$$

$$\beta_{00k} = \gamma_{000} + u_{00k} \quad (51)$$

Contextual model. Two contextual models were estimated. The first contextual model included all of the student, class, and instructor-level control variables (student gender, student age, student ethnicity, prior student interest, student effort, electivity, class size, class starting time, class meeting pattern, class length, class-level, instructor gender, instructor ethnicity, and instructor position). All of the student, class and instructor-level fixed effects were retained regardless of statistical significance because of their use as control variables.

The second contextual model included the addition of the student course rating variable as a class-level predictor of final course grade. By comparing the τ estimates between the first and second contextual model, the proportion of variance in student final course grade explained by student ratings of instructor effectiveness above and beyond what was explained by the student, class, and instructor-level control variables was calculated. This final model was used for interpreting results.

$$Y(\text{Grade})_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{StudGen})_{jk} + \pi_{2jk}(\text{StudAge})_{jk} + \pi_{3jk}(\text{StudEth})_{jk} + \pi_{4jk}(\text{Interest})_{jk} + \pi_{5jk}(\text{Electivity})_{jk} + \pi_{6jk}(\text{Eval})_{jk} + e_{ijk} \quad (52)$$

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(\text{ClassSize})_{jk} + \beta_{02k}(\text{MeetTime})_{jk} + \beta_{03k}(\text{ClassLength})_{jk} + \beta_{04k}(\text{Level})_{jk} + r_{0jk} \quad (53)$$

$$\beta_{00k} = \gamma_{000} + \gamma_{001}(\text{InstructGen})_{jk} + \gamma_{002}(\text{InstructEth})_{jk} + \gamma_{003}(\text{InstructRank})_{jk} + u_{00k} \quad (54)$$

CHAPTER IV

RESULTS

Descriptive Statistics

In this section descriptive statistics (mean, standard deviation, and frequency) are presented for the predictor and dependent variables used in this study. There is some redundancy between these results and the sample demographic variables presented in the methods chapter. This redundancy is due to the fact that the demographic variables also serve as predictors of student course rating in the second research question.

Table 5 presents the descriptive statistics for the predictor variables. The majority of students were female (73.1%) and white (84.1%). The average student age was around 28 years ($SD = 9.29$). The mean reported interest in the subject at the beginning of the course was 3.61 (on a 5 point Likert-scale) which is between average and high while the mean amount of student effort was slightly higher at 4.07 (a “high” amount of effort).

The majority of courses ($n = 475$) began between in the late afternoon, between 4:01 and 6:00 PM (52.6%). The remaining 225 courses were spread rather evenly throughout the day: (a) 8: 01 to 10 AM (18.1%), (b) 10:01 to 12 PM (5.9%), (c) 12:01 to 2PM (13.5%), (d) 2:01 to 4PM (4.4%), and (e) 6:01 to 8PM (5.5%). Average course enrollment was 19.10 and an average length of 136.7 minutes (about two and a half hours). Close to 80% of the classes met once a week. The classes were almost evenly split among the undergraduate (45.3%) and graduate (54.7%) levels.

The 178 instructors exhibited similar demographics to the student respondents, with the majority of instructors reporting that they are white (87.1%) and female (63.5%). Most of the course instructors were listed as instructor, lecturer, or staff (57.9%); nearly 19% were listed as an associate professor, almost 12% were listed as a full professor, 8% as assistant professor, and 4% were listed as a graduate assistant.

Table 5

Descriptive statistics for the predictor variables

Variable	Mean	SD	Frequency of 0 (%)	Frequency of 1 (%)
Student level				
Gender	--	--	1138 (26.9%) Male	3094 (73.1%) Female
Race	--	--	3557 (84.1%) White	675 (15.9%) Non-white
Electivity	--	--	298 (7 %) Not required	3934 (93%) Required
Age	28.36	9.49	--	--
Prior interest	3.61	1.00	--	--
Effort	4.07	0.82	--	--
Course level				
Class level	--	--		
Meeting pattern	--	--		
Enrollment	19.10	11.61	--	--
Length (minutes)	136.77	44.62	--	--
Instructor level				
Gender	--	--	65 (36.5%) Male	113 (63.5%) Female
Race	--	--	155 (87.1%) White	23 (12.9%) Non-white

The mean values for the student evaluation instrument items are presented in Table 6. Eighteen of the 19 mean values for the individual items were greater than 4, falling between “high”/”agree” and “very high”/”strongly agree”. The one item with a mean value less than 4 was the item measuring the student’s overall impression of the course. The mean value of 3.89 neared the qualitative descriptor of “above average”. These scores indicate that on average, student’s rated the various dimensions of instruction and course organization above average, with most mean values between 4 and 5 on the 5 point Likert-scale. The average sum total of the evaluation instrument was 81.21 out of a maximum value of 95. When this value is divided by the number of items ($n = 19$) the result is 4.27.

Table 6

Descriptive statistics for the outcome variables

	<i>Mean² (Mean³)</i>	<i>SD</i>
The grading system was clearly explained	4.34 (3.34)	0.95
Course goals were clear.	4.38 (3.38)	0.91
Grading in the course was based on how well students performed on assigned work.	4.47 (3.47)	0.77
There was agreement between announced goals of the course and what was actually taught.	4.36 (3.36)	0.91
Homework assignments and projects covered materials that had been presented.	4.31 (3.31)	0.90
The instructor was well-prepared for class.	4.46 (3.46)	0.84

² This value was calculated with the variable coded from 1 to 5.

³ This value was calculated with the variable centered at 0 for HLM analysis, coded from 0 to 4.

The instructor found alternative ways of explaining material when students didn't understand.	4.28 (3.28)	0.99
Judging by presentations and answers to questions, the instructor displayed a clear understanding of the course topics.	4.56 (3.56)	0.78
Difficult concepts were explained in a helpful manner.	4.28 (3.28)	1.01
Course content was related to general knowledge and experience external to the course.	4.41 (3.41)	0.82
The overall organization of the course (relationship among lectures, readings, and classroom activities) contributed to learning.	4.28 (3.28)	0.98
How much did you learn from this course?	4.07 (3.07)	1.04
My overall impression of this course was...	3.89 (2.89)	1.07
The instructor's teaching was...	4.07 (3.07)	1.03
Class presentations were intellectually stimulating.	4.14 (3.14)	1.10
The instructor caused me to think critically.	4.28 (3.28)	0.99
In this class, standards for student performance were...	4.12 (3.12)	0.81
Textbooks and other helped in learning the course content.	4.31 (3.31)	0.90
Textbooks and other materials fit the goals of the course.	4.28 (3.28)	0.98
Sum of total evaluation	81.21 (62.21)	14.32
Sum of factor one (11 items)	48.13 (37.13)	8.38
Sum of factor two (6 items)	24.57 (18.57)	5.20
Sum of factor three (2 items)	8.51 (6.51)	1.83

Research Question One

The first research question “do the scores obtained in the study exhibit adequate reliability and construct validity?” was answered by examining the validity and reliability

of the obtained student ratings. Results of the factor analysis and reliability analysis are presented in the following section.

Construct Validity

The construct validity of the student course ratings was assessed by conducting a factor analysis. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was .96, exceeding the commonly used cutoff value of .50 (Kaiser, 1970). These results indicated that there was common variance among the nineteen items and that the data were appropriate for a factor analysis (Stevens, 2001). Results of Bartlett's Test of Sphericity were statistically significant at the .05 alpha level ($\chi^2 (171) = 95002.61, p < .001$), thus rejecting the null hypothesis that the correlation matrix is an identity matrix. The variables in this study were correlated with one another.

Results of the parallel analysis and an examination of the scree plot suggested a three-factor solution. One thousand random matrices that parallel the parameters of the actual data ($N_{\text{cases}} = 5,000$; $N_{\text{variables}} = 19$) were created using SPSS syntax. The resultant mean and 95th percentile eigenvalues were compared to the initial eigenvalues from principal axis factoring of the actual dataset in order to determine the number of factors to retain. Those initial eigenvalues that exceeded the mean and 95th percentile eigenvalues of the randomly generated data were retained. The initial eigenvalues for factors one, two, and three exceed their respective randomly generated mean and 95th percentile eigenvalues, thus providing evidence for the retention of three factors (see *Appendix B*).

Similar evidence for the retention of three factors can be found in an examination of the scree plot (see *Appendix C*) which plots the nineteen initial eigenvalues in rank

order from largest to smallest from left to right. To determine the number of factors to retain one looks for the “elbow” in the plot, that point at which the amount of variance explained by each additional component is minimal. The “elbow” in this scree plot appears to be at the third factor. Given the subjective nature of evaluating the scree plot, the scree plot was used to simply confirm the results of the parallel analysis. It may also be noted that using the traditional Kaiser’s eigenvalue-greater-than-1 (K1) rule, one would retain three factors.

The three factors accounted for 63.08%, 5.09%, and 4.15% (prerotation) of the variance which combined to account for 72% of the total variance. Results in the factor correlation matrix show that factors one and two are highly correlated ($r = .803$) with much lower correlation values between factors one and three ($r = .597$) and two and three ($r = .563$).

Following the extraction of factors, oblique rotation methods were used to increase the interpretability of the results. Promax rotation with Kaiser normalization was selected because of the expected moderate correlations between the items on the student evaluation instrument and oblique rotations ability to allow items to be correlated with one another. Items with a minimum factor pattern coefficient of .40 were considered appropriate for inclusion in the rotated factors (Stevens, 2001). Using a criterion of .40 as a cutoff point, all 19 of the items yielded salient pattern coefficients on one or more factors. Factor one had 11 items with structure coefficients greater than .40, factor two had 8 items with structure coefficients greater than .40, and factor three had two items with structure coefficients greater than .40. The items “the instructor found alternative ways of explaining material when students didn’t understand” and “difficult concepts

were explained in a helpful manner “ loaded on factors one and two with a pattern coefficient greater than .40. These items were retained on factor one because it exhibited a higher pattern coefficient and fit better with the theoretical composition of factor one. (see Table 7).

Table 7.

Pattern Coefficients, Structure Coefficients, and Communalities

Student course evaluation item	Pattern coefficients (structure coefficients)			h^2
	F1	F2	F3	
The grading system was clearly explained	0.957 (0.857)	-0.142 (0.639)	0.024 (0.515)	0.741
Course goals were clear.	0.918 (0.882)	-0.033 (0.695)	-0.015 (0.514)	0.779
Grading in the course was based on how well students performed on assigned work.	0.904 (0.836)	-0.116 (0.633)	0.043 (0.516)	0.704
There was agreement between announced goals of the course and what was actually taught.	0.861 (0.894)	0.043 (0.733)	-0.002 (0.536)	0.801
Homework assignments and projects covered materials that had been presented.	0.771 (0.800)	-0.033 (0.638)	0.093 (0.534)	0.645
The instructor was well-prepared for class.	0.584 (0.816)	0.335 (0.769)	-0.062 (0.475)	0.704
The instructor found alternative ways of explaining material when students didn't understand. ⁴	0.519 (0.834)	0.447 (0.823)	-0.073 (0.488)	0.765
Judging by presentations and answers to questions, the instructor displayed a clear understanding of the course topics.	0.508 (0.778)	0.364 (0.751)	-0.037 (0.471)	0.652
Difficult concepts were explained in a helpful manner. ⁵	0.495 (0.837)	0.479 (0.836)	-0.070 (0.494)	0.780
Course content was related to general knowledge and experience external to the course.	0.472 (0.713)	0.258 (0.669)	0.058 (0.484)	0.537

⁴ Pattern coefficient was greater than .40 on factors one (.519) and two (.447).

⁵ Pattern coefficient was greater than .40 on factors one (.495) and two (.479).

The overall organization of the course (relationship among lectures, readings, and classroom activities) contributed to learning.	0.410 (0.822)	0.326 (0.797)	0.251 (0.680)	0.768
How much did you learn from this course?	-0.093 (0.663)	0.881 (0.853)	0.082 (0.522)	0.732
My overall impression of this course was...	-0.007 (0.703)	0.871 (0.875)	0.018 (0.504)	0.766
The instructor's teaching was...	0.089 (0.735)	0.859 (0.889)	-0.074 (0.462)	0.795
Class presentations were intellectually stimulating.	0.185 (0.757)	0.703 (0.859)	0.012 (0.518)	0.750
The instructor caused me to think critically.	0.224 (0.774)	0.681 (0.864)	0.007 (0.523)	0.765
In this class, standards for student performance were...	-0.193 (0.395)	0.677 (0.564)	0.075 (0.341)	0.331
Textbooks and other helped in learning the course content.	-0.036 (0.567)	0.061 (0.555)	0.928 (0.941)	0.887
Textbooks and other materials fit the goals of the course.	0.107 (0.614)	0.004 (0.565)	0.846 (0.911)	0.838
Eigenvalues	12.25	1.15	1.09	
% of variance after rotation	63.08%	5.09%	4.15%	

Factor one was composed of 11 items, eight of which assess the course organization and goals and three of which assess the examinations and grading. This factor is similar to the organization and presentation skills factor presented by Patrick and Smart (1998). Patrick and Smart's (1998) eight-item organization and presentation skills factor included the items: (a) "the teacher was well-prepared", (b) "the teacher was well organized", and (c) "the teacher made the aims of each lesson clear". These items are very similar to the following course organization items found in factor one in the present study: (a) "the instructor was well-prepared for class", (b) "the overall organization of the

course contributed to learning”, and (c) “there was agreement between announced goals of the course and what was actually taught”.

The second factor was composed of six items that assessed student learning and instructor teaching. This factor is similar to Marsh’s (1983, 1984, 1987) five-item learning/value factor which includes items such as: (a) “you found the course intellectually challenging and stimulating”, and (b) “your overall course rating”. The following similarly worded items are found in factor two in the current study: (a) “class presentations were intellectually stimulating”, and (b) “my overall impression of this course was...”.

The third factor was composed of two items that assess the textbooks and course materials: (a) “textbooks and other materials helped in learning the course content”, and (b) “textbooks and other materials fit the goals of the course”. The decision was made to retain the third factor because the items do seem to assess a unique aspect of the course organization not found in the other factors, the textbooks and course materials. A more in-depth discussion of the emergent factors and their connection to previously published factor structures can be found in Chapter IV.

Reliability

In the current sample ($n = 4780$ with complete evaluation data), the Cronbach’s alpha for all 19 evaluation items was .968 (Table 8) with a mean inter-item correlation of .617, with correlations ranging from .314 to .893. The scale mean for the 19 items is 81.54 ($SD = 14.0$) out of a maximum scale value of 95. The reliability of the 19 evaluation items calculated from the current sample (.97) is slightly higher than the previously reported value from the same college in 1988 (.94).

The first factor exhibited a Cronbach's alpha of .96, with a mean inter-item correlation of .689, with correlations ranging from .568 to .893. The scale mean for the 11 items is 48.28 ($SD = 8.18$) out of a maximum scale value of 55. When the mean value was divided by the number of items ($n = 11$), the result is 4.38 (on a 5 point scale), midway between "agree" and "strongly agree". The second factor exhibited a Cronbach's alpha of .927, with a mean inter-item correlation of .672, with correlations ranging from .456 to .853. The scale mean for the 6 items is 24.67 ($SD = 5.13$) out of a maximum scale value of 30. When the mean value was divided by the number of items ($n = 6$), the result is 4.11 (on a 5 point scale). The third factor exhibited a Cronbach's alpha of .793, with an inter-item correlation of .657. The scale mean for the 2 items is 8.50 ($SD = 1.78$) out of a maximum scale value of 10. When the mean value was divided by the number of items ($n = 2$), the result is 4.25 (on a 5 point scale). All of the reported reliability coefficients exceed the minimum reliability of .70 suggested by Nunnally (1978), with only the reliability for the textbooks/course materials factor falling below .90.

Table 8

Reliability statistics for obtained scores

Factor	Cronbach's alpha	Cronbach's alpha based on standardized items	Number of items
Factor one	.960	.961	11
Factor two	.927	.925	6
Factor three	.793	.793	2
Total course rating	.968	.968	19

Research Question Two

The second research question, “what student, course and instructor-level variables are statistically significant predictors of student ratings of instructor effectiveness?” was answered by creating a prediction model that regressed student, course, and instructor level variables on student course ratings. The relationships were analyzed with hierarchical linear modeling.

Due to the three factor structure that emerged from the factor analysis of the student course evaluation instrument, four separate HLM analyses were conducted to answer the second research question. The first model incorporated all 10 evaluation items, and the sum of the 19 items was regressed on the student, class, and instructor level predictor variables. Three additional analyses were conducted to assess the relationship between the predictor variables and the three emergent student evaluation instrument factors. The results of the HLM analyses are presented in the following section of the paper.

Overall Student Course Rating

Unconditional model. The estimated unconditional model provided the grand mean of the overall student course rating and a measure of variance at the student, course, and instructor levels (see *Appendix E*). The mean value for the overall student course rating was 62.27 (γ_{00}), out of a total possible value of 76, with a standard error of 0.61. The variability in overall student course ratings was partitioned into the variance between students within classes ($\sigma^2 / \sigma^2 + \tau_\pi + \tau_\beta$), variance between classes within instructors ($\tau_\pi / \sigma^2 + \tau_\pi + \tau_\beta$), and variance between instructors ($\tau_\beta / \sigma^2 + \tau_\pi + \tau_\beta$; Raudenbush & Bryk, 2002). The estimated variance between students within classes was .6651, meaning that the majority (66.5%) of the variability in overall course rating is due to differences between

students within classes. Most of the remaining variance was estimated at the between instructor level (.2350), with about 10% of the variance (.0999) at the between classes within instructor level.

Student level model. A random coefficients model was estimated in which overall student course rating was considered a function of student age, student ethnicity, student gender, course electivity, prior student interest, and amount of effort. Prior interest in the subject and effort were the only statistically significant predictors of overall student course rating ($p < .05$; see *Appendix F*). The between-student within-class variability was reduced from 130.53 in the unconditional model to 108.84 in the student level model, indicating that almost 17% of the variability in overall student course rating was explained by student age, student ethnicity, student gender, course electivity, prior student interest, and amount of effort.

Class level model. A contextual model was estimated in which the six student-level predictors were considered a function of the time of the class meeting, number of class meeting per week, length of the class, enrollment, and level. As shown in *Appendix G*, none of the class level predictors were statistically significant at the .05 alpha level. These results were not surprising given that only about 10% of the variability in overall course rating was due to differences between students within classes.

Instructor level model. A full model was estimated which included the instructor level predictors of gender, race, and position. The instructor-level predictor of race nearly achieved statistical significance at the .05 alpha level ($p = .054$; *Appendix H*). The student-level predictor of age achieved statistical significance in the final model ($p =$

.044), with a .06 point increase in overall student rating for each year older than the grand mean (28 years), holding all other student, class and instructor-level variables constant.

Similar predictions can be made with the statistically significant predictors from the student and course-level models. For example, holding all other variables constant, a one unit increase in prior course interest would result in a 1.16 point increase in overall student rating. Holding all other variables constant, a student that indicated “very low” interest in the course subject at the beginning of the semester would be expected to have an overall course rating of 78.18 (out of 95) while a student indicating “very high” interest would be expected to have an overall course rating of 82.82. An even more pronounced effect can be found when predicting values based on the amount of student effort as a student indicating “very low” effort would be expected to have a predicted overall course rating of 65.55, holding all other variables constant, while a student indicating “very high” amount of effort would achieve a predicted overall course rating of 85.95.

Factors as Outcome Variables

The same four-step model building process described above was employed using each of the three factors that emerged from the factor analysis as dependent variables: (a) factor one (11 items), (b) factor two (6 items), and (c) factor three (2 items). Similar results to those presented for the overall course rating as the dependent variable were found across the three factors. The following section provides an overview of the findings, highlighting similarities and differences between the models. More detailed results can be found in the appendixes.

Estimated variance. The estimated variance at the student, class, and instructor-level, gathered from the unconditional models for each of the three factors tell a similar story to that told by the overall course rating model (Table 9). Most of the estimated variability in the outcome variable (factors one, two, and three) is due to differences between students within classes. This explains why most of the statistically significant predictors of each factor came from the student-level.

Table 9.

Estimated variance at the student, course, and instructor level

Outcome variable	Estimated variance at each level		
	Student	Course	Instructor
Factor one	.6739	.0927	.2334
Factor two	.6760	.0955	.2285
Factor three	.7764	.0906	.1329
Total course rating	.6651	.0999	.2350

Factor one. Results of the random coefficients model revealed that prior student interest and effort were statistically significant predictors of factor one (see *Appendix J*). The combination of the two significant predictors, student age, student gender, student race, and electivity explained about 11% of the variance in the organization and presentation skills factor. The estimation of the course level model did not reveal any statistically significant class level predictors but in combination the predictors did account for 11% of the variance in factor one, above and beyond what was accounted for by the student-level predictors (see *Appendix K*). The addition of instructor-level

predictors in the full model did not reveal any statistically significant predictors, though instructor race did approach statistical significance with a p -value of .052. (see *Appendix L*). The non-significant instructor level predictors of race, gender, and position did account for about 4% of the variance in factor one, above and beyond what was accounted for by the student and class-level predictors.

Factor two. Similar to all of the other models, prior student interest and effort were statistically significant student level predictors. In this model, student age was a statistically significant predictor of factor two, while student race nearly reached statistical significance ($p = .054$; see *Appendix N*). The student-level predictors explained about 24% of the variance in factor two. The estimation of the course level model did not reveal any statistically significant class level predictors but course enrollment nearly reached statistical significance ($p = .055$; see *Appendix O*). The addition of instructor-level predictors in the full model did not reveal any statistically significant predictors, though similar to the previous models, instructor race did approach statistical significance with a p -value of .052. (see *Appendix P*). Additionally, the course-level class enrollment variable did achieve statistical significance. For every additional student enrolled in a class above the grand mean (24), student scores on factor two decreased by .03 points when holding all other variables constant.

Factor three. In the random coefficients model, student interest, effort, and student race were statistically significant predictors of factor three (see *Appendix R*). Holding all other variables constant, it was predicted that non-white students would score 8.68 (out of 10) on factor two compared to 8.51 for white students.

The estimation of the course level model did not reveal any statistically significant class level predictors (see *Appendix S*). The addition of instructor-level predictors in the full model did not reveal any statistically significant instructor-level predictors, though the course-level meeting pattern variable did reach statistical significance ($p = .05$; see *Appendix T*).

Unique to this model is the statistical significance of the course meeting pattern variable. Though a statistically significant predictor of the course materials and textbook rating items, course meeting pattern had a minimal impact on the predicted course rating. Holding all other variables constant, students enrolled in a course that met more than 1 day per week were predicted to have a factor three score of 8.96 (out of 10) as compared to those enrolled in a course meeting once a week with a mean score of 8.51.

Summary. The regression models using the three factors as outcome variables exhibited similar results to the total course rating model with a few differences. Because most of the variance in student course evaluations was estimated at the student-level it is not surprising that the student level predictors explained more variance in the outcome than the course and instructor-level predictors (see Table 9). The student-level variables prior student interest and effort were statistically significant predictors in every model. Student age was a significant predictor of overall course rating and the six-item second factor. Student race was a statistically significant predictor of the two-item third factor and approached significance in the factor two model.

There were only two instances amongst the four HLM models of a statistically significant course-level predictor: (a) class size was a significant predictor of factor two, and (b) class meeting pattern was a statistically significant predictor of factor three.

None of the instructor-level variables significantly predicted the outcome variable.

Instructor race did approach statistical significance in predicting the overall course rating ($p = .054$), factor one ($p = .051$) and factor two ($p = .052$).

Table 10

Summary of the statistically significant predictors of student course rating

Statistically significant predictors in full model ⁶			
Outcome variable	Student	Course	Instructor
Factor one	Prior interest, effort		
Factor two	Prior interest, effort, age	Enrollment	
Factor three	Prior interest, effort, race	Meeting pattern	
Total course rating	Prior interest, effort, age		

Research Question Three.

A three-step HLM model was estimated to assess the relationship between a student's rating of instructor effectiveness and a student's final course grade. The unconditional model provided the grand course grade mean as well as estimates of the variance in grade at the student, class, and instructor-level. A second model was estimated that included all of the student, class and instructor-level control variables. A third model included the addition of total course rating as a student-level predictor of the student's grade.

Unconditional model. The mean value for student grade was 3.77 (γ_{00}), out of a total possible value of 4.0, with a standard error of 0.02 (See *Appendix U*). This is approximately a grade of A-. The variability in student grade was partitioned into the

⁶ $p < .05$

variance between students within classes ($\sigma^2 / \sigma^2 + \tau_\pi + \tau_\beta$), variance between classes within instructors ($\tau_\pi / \sigma^2 + \tau_\pi + \tau_\beta$), and variance between instructors ($\tau_\beta / \sigma^2 + \tau_\pi + \tau_\beta$; Raudenbush & Bryk, 2002). The estimated variance between students within classes was .7169, meaning that the majority (71.62%) of the variability in student grade is due to differences between-students within- classes. The remaining variance was estimated at the between-class within-instructor (.1126) and the between-instructor (.1712) levels.

Control model. A second model was estimated that included student (prior interest, electivity, effort, gender, race, and age), class (size, level, length in minutes, meeting pattern, and starting time), and instructor-level (race, gender, position) control variables. The fixed effects of the control variables are not of interest to this research question as there is not empirical literature-based evidence to support the relationship between these variables and student grade.

Student rating as a predictor model. The final estimated model included overall student course rating as a student-level predictor of student grade (see *Appendix W*). Total evaluation rating was a statistically significant predictor of student course grade ($p = .001$). The fixed effect of .002 can be interpreted as meaning that for every 1 point decrease in total evaluation score from the grand mean of 81.21 (out of the 95 point scale) the students final course grade would decrease by .002 points. While the effect is statistically significant, the size is not substantial enough to have practical meaning. For example if a student rated a course “very high”/”very much”/”strongly agree” for every evaluation item, it is predicted that the student would earn a grade of 3.81, holding all control variables constant. If the same student rated a course “very low”/”very

little”/”strongly disagree” on all 19 evaluation items, it is predicted that the student would earn a grade of 3.66, holding all of the control variables constant.

CHAPTER V

DISCUSSION

In this chapter, a review of the major findings of this study, a discussion of the results and their relation to the existing literature, recommendations for future research, and limitations are presented.

Review of the Results

Research question one

The results from research question one (do the scores obtained in the study exhibit adequate reliability and construct validity?) suggest that the obtained scores are reliable and suggest a three factor structure for the student evaluation instrument. Results of parallel analysis and an examination of the scree plot that resulted from the factor analysis suggested the retention of three factors. Those three factors accounted for 63.08%, 5.09%, and 4.15% (prerotation) of the variance which combined to account for 72% of the total variance.

Factor one (organization and presentation skills) is composed of eleven items, eight of which assess the course organization and goals and three of which assess the presentation skills of the instructor. The second factor (learning/ability to challenge students) is composed of six items that assess the student learning and the instructor's ability to challenge students. The third factor (textbooks/course materials) is composed of two items that assess the textbooks and course materials.

Reliability estimates were calculated and reported for the overall instrument and the three factors. The Cronbach's alpha from the entire 19 items instrument was .97. The organization and presentation skills factor (11 items) exhibited a Cronbach's alpha of .96, the learning/ability to challenge students factor (6 items) exhibited a Cronbach's alpha of .93, and the textbooks/course materials factor (2 items) exhibited a Cronbach's alpha of .79.

Research question two

Due to the three factor structure that emerged from the factor analysis of the student course evaluation instrument, four separate HLM analyses were conducted (overall instrument and the three factors) to answer the second research question (what student, course and instructor-level variables are statistically significant predictors of student ratings of instructor effectiveness?). Table 10 (on page 84) displays the significant predictors for each of the four HLM models. The only variables that were statistically significant predictors in all four models were the student-level predictors of effort and prior interest in the subject. Student age was a statistically significant predictor of overall course rating while both student age and student race were statistically significant predictors of factor three. There were only two instances of statistically significant course-level predictor variables. Class size was a statistically significant predictor of factor two while class meeting pattern was a significant predictor of factor three. None of the instructor-level variables were statistically significant predictors. These findings can be explained by the fact that the majority of the variance in the outcome variables is due to differences between students within classes (see Table 9 on page 81).

Research question three

A three-step HLM model was estimated to answer research question three (do a student's ratings of instructor effectiveness predict a student's final course grade?). After creating the unconditional model, a second model was estimated that included all of the student, class and instructor-level control variables. The final model included the addition of total course rating as a student-level predictor of the student's grade.

In the final model, total evaluation score was a statistically significant predictor of student course grade ($p = .001$). The fixed effect of .002 can be interpreted as meaning that for every 1 point decrease in total evaluation score from the grand mean of 81.21 (out of 95 point scale) the students final course grade would decrease by .002 points.

Emergent Factors

The three factors that emerged from a factor analysis of the student course ratings were similar in content and composition to previously published work. Factor one in the present study was composed of 11 student course evaluation items that asked the respondent to assess the organization of the course, the instructor's preparedness, course goals, and grading. This factor is similar to several prominent factors that have been published in the student course evaluation literature. Factor one is most closely aligned with the organization and presentation skills factor presented by Patrick and Smart (1998), who conducted a factor analysis of the scores obtained from a 72 item qualitatively derived meta-inventory of instructor effectiveness in a sample of 266 Australian undergraduate psychology students. Patrick and Smart's (1998) eight-item organization and presentation skills factor included the items: (a) "the teacher was well-prepared", (b) "the teacher was well organized", and (c) "the teacher made the aims of

each lesson clear”. These items are very similar to the following course organization items found in factor one in the present study: (a) “the instructor was well-prepared for class”, (b) “the overall organization of the course contributed to learning”, and (c) “there was agreement between announced goals of the course and what was actually taught”. Course organization factors are common in the published factor analyses of student evaluation instruments. Marsh (1983, 1984, 1987) proposed a four-item organization factor in his factor analysis of the Students’ Evaluations of Educational Quality (*SEEQ*). Marsh’s organization factor included the items: (a) instructor explanations were clear, (b) course materials were prepared and clear, and (c) objectives were stated and pursued. Further evidence of an organization factor can be found in Feldman’s (1976, 1984) system for categorizing items from course evaluation instruments which included both a “preparation and organization of the course” and “clarity of course objectives” factor.

Three of the 11 factor one items assessed the grading system with the items: (a) the grading system was clearly explained, (b) grading in the course was based on how well students performed on assigned work, and (c) homework assignments and projects covered material that had been presented. Marsh (1983, 1984, 1987) posited a similarly composed three-item examinations/grading factor in his analysis of the *SEEQ*. Additional evidence of a grading/student assessment factor can be found in Feldman’s (1976, 1984) fairness of evaluation factor. The published evidence reported in the preceding paragraphs support the consideration of factor one as a course organization and grading factor, combining published organization and grading factors from several prominent factor analyses of student course evaluations.

The second factor in the present study is composed of six items that assessed student learning and instructor teaching. This factor is similar to Marsh's (1983, 1984, 1987) five-item learning/value factor which includes items such as: (a) "you found the course intellectually challenging and stimulating", and (b) "your overall course rating". The following similarly worded items are found in factor two in the current study: (a) "class presentations were intellectually stimulating", and (b) "my overall impression of this course was...". There is also evidence of similarity between factor two and Patrick and Smart's (1998) eight-item challenge factor which included the items: (a) the teacher really challenged you, and (b) the teacher had the ability to motivate you to do your best. Additional evidence of a student learning and instructor teaching factor can be found in Centra's (1994) list of commonly identified student evaluation factors which included a teacher-student interaction or rapport factor. Based on the previously reviewed literature, factor two can be considered a learning and ability to challenge factor.

The third factor in this study is composed of two items that assess the textbooks and course materials: (a) "textbooks and other materials helped in learning the course content", and (b) "textbooks and other materials fit the goals of the course". This factor is limited in scope, as it focuses exclusively on the course materials but tends to be theoretically related to the course organization, a factor common in many studies of instructor effectiveness (Braskamp & Ory, 1994; Centra, 1993, Feldman, 1988, Patrick & Smart, 1998; Sherman et. al., 1987). The decision was made to retain the third factor because the items do seem to assess a unique aspect of the course organization not found in the other factors, the textbooks and course materials. Additional evidence of a course

materials factor can be found in Feldman's (1976, 1984) relevance and value of course materials factor.

Results of the exploratory factor analysis of the student course rating scores suggested a three factor solution. Based on strong connections to previously published student evaluation factor structures, the three emergent factors can be considered (a) the course organization and grading factor (11 items), (b) the learning and ability to challenge students factor (6 items), and (c) the course materials factor (2 items).

Predictors of Student Course Ratings

Student-level predictors.

Results of the hierarchical linear modeling analyses revealed that the majority of the variance in the outcome variables is due to differences between students within classes. Given these findings, it is not surprising that most of the statistically significant predictors of student course ratings were student-level variables. Prior interest in the subject and the amount of effort put into the course were statistically significant predictors in each of the four HLM analyses (overall course rating as the outcome variable, and the three factors as outcomes). There is an abundance of empirical evidence supporting the positive relationship between prior student interest and student course ratings (Feldman, 1978; Marsh & Cooper, 1981; Prave & Baril, 1993; Wachtel, 1998). Feldman (1978) cited five studies that explored this relationship between intrinsic student interest in a course subject and student ratings. Feldman found that there was a small positive statistically significant relationship between student interest and course ratings, with correlations in the .10s to .20s. Similar results were reported in this study with zero-

order correlations between prior student interest and the outcome variables ranging from .13 to .23.

In this study, student interest was relatively high, with nearly 70% of respondents indicating that their interest in the course at the beginning of the semester was average or high with a mean item response of 2.61 (SD = 1) out of 4. Prave and Baril (1993) warn that it may be necessary to control for this potential bias in general education classes like English or mathematics, which are required for all students and may be of little interest to some students. In the current study, all of the evaluated courses took place in the college of education, meaning that students would not be taking the courses as general education requirements but taking them as part of their major/minor study concentration.

Student effort was found to be positively related to course rating in the four HLM analyses, with zero-order correlations between effort and the outcome variables ranging from .27 to .44. In the analysis of the overall course rating, holding all other predictors constant, a 1 unit increase in the amount of reported effort above the mean (3.04) would result in an increase of 5.1 in the overall course rating. Prior literature suggests a similar positive relationship between student effort and student course rating (Heckert, Latier, Ringwald-Burton, Drazeen, 2006;). Heckert et al. (2006) reported a statistically significant correlation of 0.36 between reported student effort and overall course rating in a sample of 463 college students.

There is additional supporting evidence in the literature examining the relationship between the similar construct of course workload/rigor and student course rating (Cashin & Slawson, 1977; Marsh & Overall, 1979; Marsh, 1982; Marsh, 1987). Prior reported positive correlations between workload and rating ranged from .14 to .29

(Cashin & Slawson, 1977). There is some evidence of a negative relationship between workload and course rating but the correlations were either not reported or extremely small ($r = .02$; Pohlman, 1975).

Student age was a statistically significant predictor of factor two ($r = .07$) and overall course rating ($r = .04$). In the overall course rating model, students received a minimal .06 increase in overall student rating for each year older than the grand mean (28 years), holding all other variables constant. This small positive correlation between student age and student rating has been reported in one of the few studies published that has explored this relationship ($r = .17$; Radmacher & Martin, 2001). Because instructor age was not a variable available in this study it was not possible to assess the interaction between student and instructor age. For example, it is possible that older students (age > 28 years) rate instructors higher than younger students because the instructor more closely approximates their own age. Further investigation needs to be given to this relationship before a more definitive statement about the nature of the relationship can be made.

Student race was a statistically significant predictor of the two-item third factor ($p = .50$) with a minimal predicted impact on the outcome variable. Holding all other variables constant, nonwhite students scored 0.17 points higher on the third factor than white students. Previous published research on race has focused on the relationship between instructor race and student course rating as well as the interaction between student and instructor race and student course rating. Neither instructor race nor the interaction between the race variables were significantly related to the third factor.

Course-level predictors.

Given the fact that that only about 10% of the variability in the outcome variables was due to differences between students within classes it is not surprising that most of the class-level variables were not statistically significant predictors. The only statistically significant course-level predictors were class size in the second factor as an outcome model and class meeting pattern in the third factor as an outcome model. There is empirical evidence that suggests a negative relationship between class size and rating, with smaller classes to receiving higher course ratings than large classes (Feldman, 1984; Neumann, 2000; Davies, Hirschberg, Lye, Johnston, & McDonald, 2007). Other researchers have posited a curvilinear relationship, characterized as U-shaped, with small and large classes receiving more favorable ratings than medium-sized classes (Centra & Creech, 1976; Gage, 1961; Pohlman, 1975).

In this study there was a negative statistically significant relationship between class size and the second factor. Holding all other variables constant, a one student increase in class size, above the mean (24 students), results in a .025 decrease in the second factor score. For example, holding all other variables constant, a class with an enrollment of 10 students can be expected to have a score of 24.22 (out of 30) while a class with 100 students can be expected to have a score of 22.67 on the second factor. The second factor included items related to instructor teaching and student learning, thus it would make sense that class size would share a relationship with this factor. The delivery of course instruction on the other hand can be expected to vary significantly in a class with 5 students and a lecture-style course with 100 students. Factors one and three included items related to the organization and structure of the course, things that would not be expected to vary based on the size of the course. The grading system or selection

of textbook would not be expected to vary between classes with an enrollment of 10, 30 or 100 students.

There was also a statistically significant relationship between class meeting pattern and the third factor. Holding all other variables constant, classes that met more than one day per week scored almost half of a point (0.45) higher on the third factor than those courses that met one day per week. While this finding is significant there is no other evidence of the existence of such a relationship in the other factors nor is there any evidence of a significant relationship between student course rating and the other class schedule variables (class meeting time and class length). There is evidence of a similar relationship between class meeting pattern and rating in a study conducted by Reardon, Payan, Miller, and Alexander (2008). Despite the significance of these results, it is difficult to generalize these findings because of the lack of replication and scarce literature on this topic.

The non-significant relationship between class meeting time and course rating supports the limited published studies of the relationship between the two variables. Feldman (1978) examined results from 11 studies and found a non-statistically significant relationship between class meeting time and student rating in seven of the 11 (63.6%) studies. While Yongkittikul, Gillmore, and Brandenburg (1974) did report significant group differences on course evaluations across nine course times, the effect size was minimal ($\omega^2 = .01$). Similarly there is limited published research on the relationship between class length/meeting pattern and course rating. Due to the limited published studies that have explored these relationships it is difficult to make a definitive statement about the existence or nonexistence of a relationship. The non-significant findings in the

present study should serve as additional data for future meta-analyses exploring the impact of the class schedule on student course ratings.

Instructor-level predictors.

None of the instructor-level predictors were statistically significant predictors of student course rating. It is not surprising that instructor experience (measured in this study by the instructor's position) failed to significantly predict the outcome variables given that most of the extensive published research on the relationship reports similar findings (Aleamoni, 1976; Brandeburg, Slinde, & Batista, 1977; Centra & Creech, 1976; Feldman, 1983). Feldman (1983) identified 33 studies that explored the relationship between instructor rank and course rating and reported no relationship in two-thirds of the studies ($n = 22$).

Similarly the non-significant relationships and interaction effects between student and instructor gender and rating and student and instructor race and student course rating were expected based on the limited body of existing research (Feldman 1992, 1993; Ludwig & Meacham, 1997; Wachtel, 1998). As stated previously, because of the limited published studies that have explored these relationships it is difficult to make a definitive statement about the existence or nonexistence of a relationship. The non-significant findings in the present study should serve as additional data for future meta-analyses exploring the impact of the student and instructor gender and race on student course ratings.

Course rating as a predictor of final course grade

Perhaps the most publicized relationship found in a review of the course evaluation literature is the potential relationship between student expected course grade

and rating. There exists evidence in several studies that posit a positive correlation between student's expected course grade and evaluation of instructor effectiveness (Braskamp & Ory, 1994; d'Appolonia & Abrami, 2007; Marsh, 1987; Marsh & Dunkin, 1992; McPheson & Jewel, 2007). In a review of the literature published on this relationship between 1924 and 1998, Aleamoni (1999) reported that 37 of 62 cited studies found a statistically significant positive relationship between actual or expected course grade and student ratings of instructor effectiveness. Twenty-four studies reported no relationship between grade and student rating, and one study reported a statistically significant negative correlation. The mean correlation across the 62 studies was 0.18 ($Mdn = 0.14$, $SD = 0.16$) which is indicative of a relatively weak relationship. Other meta-analytic studies have cited a mean correlation between .10 and .30 (Centra, 2003; Feldman, 1997), a relationship of moderate magnitude.

In this study it was not possible to obtain a measure of the student's expected course grade. In place of expected course grade, the student's final course grade was collected. Because the act of assigning a course evaluation rating precedes the distribution of final course grade, course rating was treated as a predictor of grade in the HLM analysis. After controlling for all of the student, class, and instructor-level variables, overall course rating was a statistically significant predictor of final student course grade ($p = .001$). The fixed effect of .002 can be interpreted as meaning that for every 1 point decrease in total evaluation score from the grand mean of 82.21 (out of 95 point scale) the students final course grade would decrease by .002 points. While the effect is statistically significant, it does not have a very large impact on the predicted student grade. A measure of the proportion reduction of variance at the student level

shows that the addition of total course rating explained a miniscule .2% of variance in student grade above and beyond what had been accounted for by the student-level control variables. The correlation between overall course rating was statistically significant ($p < .001$) at .09 but weaker in magnitude than the values reported in previous studies.

Limitations

As with any research, there are imperfections and things that the researcher feels could be done to improve the study. For example, the results of this study cannot be generalized to the worldwide population of college students. The sample in the study is limited to undergraduate and graduate students enrolled in a College of Education and Human Development at a large metropolitan research university in the southern United States. As such, the sample is somewhat homogeneous in its interest in the course subjects as the courses would in most cases be considered part of an academic major and undertaken at the latter stages of undergraduate education and post-baccalaureate study in preparation for a potential career in the field of education or human development. The high level of interest in the course subject may have contributed to the limited range of student course ratings and had a dampening effect on the effect size of any statistically significant relationship. Because the study took place within the context of a College of Education and Human Development, generalization of the results to the general population of college students is limited.

In the present study, the most glaring limitation is the presence of missing data. The missingness of the data comes from two sources which result in two separate limitations: (a) incomplete student course evaluation data, and (b) missing data in the university-maintained database. Given that student course evaluations are voluntary in

nature it could be expected that the response rate would be less than 100%. An additional change in the response rate can be expected due to the fact that the university at which the research was conducted recently moved to online delivery of course evaluations. Previous research has provided contradictory statements about the impact of the move from paper-and-pencil to online course evaluations. Dommeyer, Baum, Chapman, and Hanna (2003) reported that response rates were lower (29%) in the online format than the traditional paper-and-pencil method (70%) while Chang (2004), who reported response rates of 79% for paper-and-pencil evaluations and 95.3% for online evaluations in a sample of 1,052 courses. In the current study the mean response rate for the 475 courses with complete class and instructor-level data was 55.9% ($SD = 19.5\%$) with individual course response rates ranging from 7% to 100%. Without having data from the non-respondents it is unclear how the participant and non-participant students may have differed in their assessment of the course and instructor.

One potential solution to the low response rate may be the use of incentives. There is limited evidence that the response rate for online evaluations may be higher if students are presented with incentives to complete the evaluation. The literature suggests several strategies for increasing the response rate, including: (a) instructors encouraging students to complete the evaluations, (b) providing an explanation of what the evaluation results are used for, (c) granting early access to grades for completing the evaluation, (d) providing bonus points for completing the evaluation, (e) early access to registration for evaluation completers, or (f) the use of prizes that can be won by evaluation completers (Anderson et al., 2005; Chang, 2004; Johnson, 2002). Dommeyer, Baum, Chapman, and

Hanna (2003) found that when any type of grade incentive was used, the online format was comparable to the traditional methods in terms of response rate.

Another limitation of the current study was missing predictor data in the database from which the data were collected. Of the original sample of 5,629 course evaluations from courses with five or more student respondents, 569 course evaluations were missing on the entirety of student demographic variables and were removed from the sample. The decision was made to impute predicted values for the cases with missing data on the student level predictor variables. Prior to the use of multiple imputation, several additional cuts were made to the data. There were 249 student evaluations that had incomplete grades or were graded on a pass/fail scale. Because of the interest in the relationship between final course grade and student evaluation of perceived instructor effectiveness those courses which did not grade on a 4.0 scale (e.g. pass/fail) were removed. Eighty-nine evaluations were missing on 20% ($n = 5$) or more of the course evaluations items and were not included in the imputation. A final data reduction was made to remove those cases with missing course or instructor predictor data ($n = 490$). The decision was made not to impute values for cases with missing data at the course or instructor level because of the limited number of covariates that could be used to estimate plausible values.

The reductions described above and in more depth in Chapter II, left a sample of 4,232 course evaluations, a reduction of 24.8% from the original sample. While the reductions do represent a substantial loss in data, the researcher feels that these decisions helped to maintain the integrity of the data. It would have been irresponsible to impute missing data without having a significant number of covariates. Additionally, if the

predictor data included imputed data estimated based on responses to the outcome variables, the resultant predictor data would have been improperly affected by the very outcome data it was collected to predict.

Implications

By and large, the findings reported in this study are consistent with the previous literature. Most of the variability in student course rating was estimated at the student-level and this was reflected in the fact that most of the statistically significant relationships were found at the student-level. Prior student course interest and the amount of student effort were statistically significant predictors of student course rating in all four regression models. These findings were supported by previous studies and provide further evidence of such relationships.

While there was evidence of statistically significant relationships between several of the predictor variables and student course ratings, particularly at the student-level, the magnitude of those relationships was minimal. In combination the student-level variables accounted for a respectable amount of the variance in the student course ratings (ranging from 11% to 24%) but taken individually, the significant variables did not have an overwhelming impact on student course ratings. These findings provide evidence of the validity of the student course evaluation instrument and suggest that the potentially “biasing” variables, while having statistically significant relationships with course ratings, would not have an overly concerning estimated impact on overall course ratings.

The weakness of the magnitude of the relationships may be due in part to the limited range of responses. Because the study took place in a College of Education and Human Development, the students are in the advanced stages of their education and

completing courses relevant to their academic majors and potential future careers. This phenomenon is evidenced by the high student interest in the course (*mean* of 3.61 out of 5). Additionally the average overall course rating in this study was 81.21 out of a potential score of 95, which averages out to a score of 4.27 out of 5. With mean student course ratings near the high end of the 5 point Likert-scale, there is limited variability in the scores and thus the effect size of a significant relationship would be limited. It could be hypothesized that more profound effect sizes may be found in a more heterogeneous student population, with more required and entry-level college courses.

Perhaps equally as compelling as the statistically significant results are the reported non-significant predictor variables. Several of the non-significant predictors, particularly class meeting time, length and pattern, and student and instructor race have not been studied extensively as potential predictors of course evaluation ratings. The non-significance and directionality of these relationships are supported by the limited number of existing studies. While these results do not provide conclusive evidence of the existence or nonexistence of such relationships, they do provide additional results that can be used in future meta-analyses.

With regard to the relationship between student rating and course grade, the results in this study support the existence of a weak positive relationship. Previous studies have reported a stronger relationship between grade and course evaluation but have used expected course grade as a predictor of the evaluation rating. It is interesting that when using course rating as a predictor of final course grade the magnitude of the reported relationship was much smaller. This is probably due to the fact that as a construct, student achievement in a class is affected more by effort, interest, and thinking/reasoning

abilities. It would be interesting to examine in a future study both the relationship between expected course grade and evaluation rating and evaluation rating and final course grade with the same student sample. Based upon conventional wisdom and the findings of this study, one could theorize that both relationships would be positive and weak to moderate in magnitude, with perhaps the grade as an antecedent relationship being slightly stronger.

The implications for educational practitioners vary on a local and global level. Locally, educators in the College of Education and Human Development at which the study was conducted can consider scores obtained from the evaluation instrument reliable, as evidenced by the internal consistency estimates provided in the current study as well as initial calibration of the instrument (Petrosko, 1988). The internal consistency values far exceed minimal values for reliable scores with a reported Cronbach's alpha for the overall course rating of .97 ($> .70$; Nunnally, 1978). Additionally, the three factor solution provides evidence of construct validity and the emergence structure resembles several prominent factor structures in the course evaluation literature. While, it may not be appropriate to make such a statement at the global level, the obtained scores from the College of Education and Human Development at a large urban research university were not widely affected or biased by the variables explored in this study. The statistically significant relationships were of minimal practical significance and of limited magnitude.

At the global level, the major contribution of this study is its addition to the body of literature on the relationship between student, course, and instructor-level variables and student ratings of instructor effectiveness. While the findings may not be universally generalizable, the results can be considered as additional data to be considered in

assessing the effect of biasing variables. Future researchers and meta analysts can consider the findings as additional evidence in coming to a consensus decision about the impact of the student, course, and instructor-level variables.

Future Research

Replication is essential in research and has been prevalent in the extensive body of literature on student course evaluation. Therefore, future research should be conducted exploring the relationships between the contextual variables and student course ratings utilizing a more heterogeneous college student population, not limited to students in a specific college within a university. Another potentially meaningful avenue of research may be to gauge student feelings about the utility of course evaluations. Given the low response rate in the current study, it may be worth investigating student's feelings about the usefulness of course ratings as a potential explanation for such a low response rate.

The field of research into student course evaluation is one that is without a prevalent theory but that is built on existing research. Because of the interest in the relationship between potentially biasing variables and student course ratings, many of the studies utilize regression methods and frame the studies in a similar fashion, by building prediction models from the extensive body of empirical data. A major contribution of future evaluation studies to the body of student evaluation literature would be the creation of strong conceptual framework. Such a framework could be used in combination with the vast body of literature to inform and frame future research studies.

REFERENCES

- Abrami, P.C., d'Apollonia, S., & Cohen, P. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82*, 219-231.
- Aleamoni, L.M. (1976). Typical faculty concerns about student evaluation of instruction. *National Association of College and Teachers of Agricultural Journal, 20*, 16-21.
- Aleamoni, L.M. (1981). Student ratings of instruction. In J. Millman (Ed.) *Handbook of Teacher Evaluation*. Beverly Hills, Sage Publications.
- Aleamoni, L.M. (1999). Student rating myth versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*, 153-166.
- Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley, L., Mohanty, G., & Andersen, K., & Miller, E.D. (1997). Gender and student evaluations of teaching. *Political Science and Politics, 30*, 216-219.

- Anderson, H.M., Cain, J., & Bird, E. (2005). Online students course evaluations: Review of literature and a pilot study. *American Journal of Pharmaceutical Education*, 69, 34-43.
- Bavishi, A., Madera, J.M., & Hebl, M.R. (2010). The effect of professor ethnicity and gender on student evaluations: Judged before met. *Journal of Diversity in Higher Education*, 3(4), 245-256.
- Bentler, P.M., & Kano, Y. (1990). On the equivalence of factors and components. *Multivariate Behavioral Research*, 25, 67-74.
- Birnbaum, M. H. (2000). *A survey of faculty opinions concerning student evaluation of teaching*. Retrieved October 28, 2010, from <http://psych.fullerton.edu/mbirnbaum/faculty3.htm>.
- Blunt, A. (1991). The effects of anonymity and manipulated grades on student ratings of instructors. *Community College Review*, 18, 48-54.
- Boretz, E. (2004). Grade inflation and the myth of student consumerism. *College Teaching*, 52, 42-46.
- Brandenburg, D.C., Slinde, J.A., & Batista, E.E. (1977). Student ratings of instruction: Validity and normative interpretations. *Research in Higher Education*, 7, 67-78.
- Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco: Jossey-Bass.
- Braskamp, L.A., Brandenburg, D.C., & Ory, J.C. (1984). *Evaluating Teaching Effectiveness*. Newbury Park, CA: Sage Publications.
- Brown, G., & Atkins, M. (1988). *Effective teaching in higher education*. London: Methuen.

- Canday, S.D., Mendelson, M.A., & Hardin, J.H. (1978). The effect of timing on the validity of student ratings. *Journal of Medical Education*, 53, 958-964.
- Cashin, W.E. (1988). Student ratings of teaching: A summary of the research. *IDEA paper* (Kansas State University. Center for Faculty Evaluation and Development in Higher Education), no. 25.
- Cashin, W.E. (1995). Student ratings of teachers: The research revisited. *IDEA paper* (Kansas State University. Center for Faculty Evaluation and Development in Higher Education), no. 32.
- Cashin, W.E., & Slawson, H.M. (1977). Description of database 1976-1977. *IDEA paper* (Kansas State University. Center for Faculty Evaluation and Development in Higher Education), no. 2.
- Centra, J.A. (1978). Using student assessments to improve performance and vitality. In W.R. Kirschling (Ed.) I, 31-49. San Francisco: Jossey_Bass.
- Centra, J.A. (1993). *Reflective faculty evaluation*. San Francisco: Jossey-Bass.
- Centra, J.A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, 44, 495-518.
- Centra, J.A., & Creech, J.R. (1976). *The relationship between student, teacher, and course characteristics and student ratings of teacher effectiveness*. Princeton, N.J.: Educational Testing Service.
- Centra, J.A., & Gaubatz, N.B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education*, 71, 17-33.

- Chang, T. (2004). The results of student ratings: Paper vs. online. *Journal of Taiwan Normal University: Education*, 49, 171-186.
- Cohen, E A. (1987). *A critical analysis and reanalysis of the multisection validity meta-analysis*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.
- Cooke, L. S. (1952). An analysis of certain factors which affect student attitudes toward a basic college course, effective living. Unpublished doctoral dissertation, Michigan State College.
- Costin, F. (1968). A graduate course in the teaching of psychology: Description and evaluation. *Journal of Teacher Education*, 19, 425-432.
- Costin, F., Greenough, W.T., Menges, R.J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, 41, 511-535.
- d'Apollonia, S. & Abrami, P.C. (1997). Navigating student ratings of instruction. *American Psychologist*, 52, 1198-1208.
- Davies, M., Hirschberg, J., Lye, J., Johnston, C., & McDonald, I. (2008). Systematic influences on teaching evaluations: The case for caution. *Australian Economic Papers*, 46, 18-38.
- Dommeier, C.J., Baum, P., Chapman, K.S. & Hanna, R.W. (2002). Attitudes of business faculty towards two methods of collecting teaching evaluations: Paper vs. online. *Assesment & Evaluation in Higher Education*, 27, 455-62.
- Doyle, K. O., & Crichton, L. I. (1978). Student, peer, and self-evaluations of college instructors. *Journal of Educational Psychology*, 70, 815-826.

- Feldman, K.A. (1972). Some theoretical approaches to the study of change and stability of college students. *Review of Educational Research*, 42, 1-26.
- Feldman, K.A. (1976a). Grades and college students' evaluation of their courses and teachers. *Research in Higher Education*, 4, 69-111.
- Feldman, K.A. (1976b). The superior college teacher from the students' view. *Research in Higher Education*, 5, 243-288.
- Feldman, K.A. (1977). Consistency and variability among college students in rating their teachers and courses: A review and analysis. *Research in Higher Education* 6, 223-274.
- Feldman, K.A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education*, 9, 199-242.
- Feldman, K.A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education*, 9, 199-242.
- Feldman, K.A. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education*, 10, 149-172.
- Feldman, K.A. (1983). Seniority and experience of college teachers as related to evaluations they receive. *Research in Higher Education*, 18, 3-124.
- Feldman, K.A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education*, 21, 45-116.
- Feldman, K.A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis.

Research in Higher Education, 24, 139–213.

Feldman, K.A. (1988). Effective college teaching from the students' and faculty's view:

Matched or mismatched priorities? *Research in Higher Education*, 28, 291-329.

Feldman, K.A. (1989a). Instructional effectiveness of college teachers as judged by

teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education*, 30, 137–194.

Feldman, K.A. (1989b). The association between student ratings of specific instructional

dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*, 30, 583–645.

Feldman, K.A. (1990). An afterword for “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* 31, 315–318.

Feldman, K.A. (1992). College students' views of male and female college teachers: Part

I- evidence from the social laboratory and experiments. *Research in Higher Education*, 33, 317-375.

Feldman, K.A. (1993). College students' views of male and female college teachers: Part

II- evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, 34, 151-211.

Feldman, K.A. (1997). Identifying exemplary teachers and teaching: Evidence from

student ratings. *The Scholarship of Teaching and Learning in Higher Education: An Evidence Based Perspective*, Section II, 93-143.

- Feldman, K.A. (2003). Identifying exemplary teachers and teaching: Evidence from student ratings. In R.P. Perry and J.C. Smart (eds.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, 93–143.
- Firth, M. (1979). Impact of work experience on the validity of student evaluations of teaching effectiveness. *Journal of Educational Psychology*, 71, 726-730.
- Frey, P.W. (1976). Validity of student instructional ratings as a function of their timing. *Journal of Higher Education*, 47, 327-336.
- Frey, P.W. (1976a). Validity of student instructional ratings: Knowledge, attitude and practice of users of student ratings of instruction. (AERA Paper).
- Fuhrmann, B.S., & Grasha, A.F. (1983). *A Practical Handbook for College Teachers*. Boston: Scott Foresman Co.
- Gage, N.L. (1961). The appraisal of college teaching. *Journal of Higher Education*, 32, 17-22.
- Greenwald, A.G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, 52, 1182-1886.
- Greenwald, A.G., & Gilmore, G.M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, 52, 1216-1217.
- Guthrie, E.R. (1954). *The Evaluation of Teaching: A Progress Report*. Seattle: University of Washington.
- Haskell, R.E. (1997). Academic freedom, tenure, and student evaluation of faculty: Galloping polls in the 21st century. *Education Policy Analysis*. Retrieved online at <http://www.bus.lsu.edu/accounting/faculty/lcrumbley/educpoly.htm>.

- Hayton, J.C., Allen, D.G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, 7, 191-205.
- Heckert, T.M., Latier, A., Ringwald-Burton, A., & Drazen, C. (2006). Relations among student effort, perceived class difficulty and student evaluations of teaching: Is it possible to “buy” better evaluations through lenient grading? *College Student Journal*, 40, 588-596.
- Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 32, 95-105.
- Hmieleski, K. (2000). *Barriers to online evaluation: Surveying the nation's top 200 most wired colleges*. Unpublished report. Rensselaer Polytechnic Institute. Troy, New York.
- Johnson, T. (2002). *Online student ratings: Will students respond?* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.
- Kaiser, H. (1970). A second generation little jiffy. *Psychometrika*, 35, 401-415.
- Koblitz, N. (1990). Are student ratings unfair to women? *Newsletter of the Association for Women in Mathematics*, 20, 17-19.
- Layne, B.H., DeCristofor, J.R. & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education*, 40, 221-232.
- Linn, R.L. (1968). A Monte Carlo approach to the number of factors problem. *Psychometrika*, 33, 37-71.

- Lovell, G.D., & Haner, C.F. (1955). Forced-choice applied to college faculty rating. *Educational and Psychological Measurement, 15*, 291-304.
- Ludwig, J.M., & Meacham, J.A. (1997). Teaching controversial courses: Student evaluations of instructors and content. *Educational Research Quarterly, 21*, 27-38.
- Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology, 75*, 150-166.
- Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. *Journal of Educational Psychology, 71*, 149-160.
- Marsh, H.W. & Dunkin, M. (1992). Students' evaluations of university teaching: A profile analysis. *Journal of Higher Education, 64*, 1-18.
- Marsh, H.W. & Overall, J.U. (1980). Validity of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal, 72*, 468-475.
- Marsh, H.W. & Roche, L.A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*, 1187-1197.
- Marsh, H.W. (1982). Validity of students' evaluations of college teaching: A multitrait multimethod analysis. *Journal of Educational Psychology, 74*, 264-279.

- Marsh, H.W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, 76, 707-754.
- Marsh, H.W. (1987). Students' evaluation of university teaching: Research findings, methodological issues and directions for future research. *International Journal of Educational Research*, 11, 253-388.
- Marsh, H.W., & Cooper, T.L. (1981). Prior subject interest, students' evaluations, and instructional effectiveness. *Multivariate Behavioral Research*, 16, 83-104.
- McKeachie, W.J. (1990). Research on college teaching: The historical background. *Journal of Educational Psychology*, 82, 189-200.
- McPherson, M.A., & Jewell, R.T. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly*, 88, 868-881.
- Medley, D.M. (1987). Criteria for evaluating teaching. In M.J. Dunkin (ed.), *The International Encyclopedia of Teaching and Teacher Education*, 169-190. Australia: Pergamon Press.
- Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology*, 75, 138-149.
- Neumann, R. (2000). Communicating student evaluation of teaching results: Rating Interpretation Guides (RIGs). *Assessment & Evaluation in Higher Education*, 25, 121-34.
- Ory, J.C., & Parker, S.A. (1989). Assessment activities at large, research universities. *Research in Higher Education*, 30, 375-385.

- Page, C.F. (1974). *Student evaluation of teaching: The American experience*. London: Society for Research into Higher Education.
- Paolo, A.M., Bonaminio, G.A., Gibson C., Partridge T. &, Kallail K. (2000). Response rate comparisons of e-mail and mail-distributed student evaluations. *Teaching and Learning in Medicine*, 12, 81-88.
- Patrick, J., & Smart, R. (1998). An empirical evaluation of teacher effectiveness: the emergence of three critical factors. *Assessment and Evaluation in Higher Education*, 23 (2), 165-178.
- Petrosko, J.M. (1990). *Course evaluation data for the University of Louisville School of Education*. Unpublished manuscript, Foundations of Education Department, University of Louisville, Louisville, Kentucky.
- Pohlmann, J.T. (1975). A multivariate analysis of selected class characteristic and student ratings of instruction. *Multivariate Behavioral Research*, 10, 81-91.
- Prave, R.S. & Baril, G.L. (1993). Instructor ratings: Controlling for bias from initial student interest. *Journal of Education for Business*, 68, 362-366.
- Radmacher, S.A., & Martin, D.J. (2001). Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis. *The Journal of Psychology*, 135 (3), 259-268.
- Rashdall, H. (1936). *The Universities of Europe in the Middle Ages, Volume 1*. Edited by F.M. Powicke & A.B. Emden. London: Oxford University Press.
- Reardon, J., Payan, J., Miller, C., & Alexander, J. (2008). Optimal class length in marketing undergraduate classes: An examination of preference, instructor evaluations, and student performance. *Journal of Marketing Education*, 30, 12-20.

- Russell, D.W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin*, 28, 1629-1646.
- Ryan, J.J., Anderson, J.A., & Birchler, A.B. (1980). Student evaluation: The faculty responds. *Research in Higher Education*, 12, 317-333.
- Schneeweiss, H. (1997). Factors and principal components in the near spherical case. *Multivariate Behavioral Research*, 32, 375-401.
- Seiver, D.A. (1983). Evaluations and grades: A simultaneous framework. *The Journal of Economic Education*, 14 (3) 32-38.
- Seldin, P. (1993). How colleges evaluate professors. 1983 vs. 1993. *American Association for Higher Education Bulletin*, 41 (7), 3-7.
- Seldin, P. (1999). Current practices-good and bad-nationally. *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions*. Bolton, MA: Anker.
- Shapiro, E.G. (1990). Effect of instructor and class characteristics on students' class evaluations. *Research in Higher Education*, 31, 135-148.
- Smalzreid, N.T., & Remmers, H.H. (1943). A factor analysis of the Purdue Rating Scale for Instructors. *Journal of Educational Psychology*, 34, 363-367.
- Spencer, P.A. & Flyr, M.L. (1992). The formal evaluation as an impetus to classroom change: Myth or reality? (Research/Technical Report, Riverside, CA.)
- Spooner, F. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching*, 52(4), 134- 141.

- Stalnaker, J.M., & Remmers, H.H. (1928). Can student discriminate traits associated with success in teaching? *Journal of Applied Psychology*, 12, 602-610.
- Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Applied Social Research Methods Series (Vol. 46). Thousand Oaks, CA: Sage.
- VanLeeuwen, D.M., Dormody, T.J., Seevers, B.S. (1999). Assessing the reliability of student evaluations of teaching (SETS) with generalizability theory. *Journal of Agricultural Education*, 40 (4), 1-9.
- Wachtel, H.K. (1998). Student evaluation of college teaching effectiveness: a brief review. *Assessment & Evaluation in Higher Education*, 24, 191-211.
- Webb, B. W., & Nolan, C. Y. (1955). Student, supervisor, and self-ratings of instructional proficiency. *Journal of Educational Psychology*, 46, 42-46.
- Widaman, K.F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, 28, 263-311.
- Witt, J., & Burdalski, D. (2003). Regarding the Timing of Student Course/Instructor Evaluations. *Journal of the Academy of Business Education*. Retrieved January 14, 2011 from <http://www.abe.villanova.edu/proc2003/witt.pdf>.
- Worthington, A.C. (2002). The impact of student perceptions and characteristics on teaching evaluations: A case study in finance education. *Assessment and Evaluation in Higher Education*, 27, 49-64.
- Yongkittikul, C., Gillmore, G.M., & Brandenburg, D.C. (1974). *Does the time of a course meeting affect course ratings by students?* (Research Report No. 346). Urbana,

IL: University of Illinois, Office of Instructional Resources, Measurement, and Research Division.

Appendix A

Student Course Evaluation Instrument

1. Is this course required for your program?
 - a. Yes
 - b. No
 - c. Not applicable or cannot answer
2. At the beginning of the course, what was your interest in the course content?
 - a. Very low
 - b. Low
 - c. Average
 - d. High
 - e. Very high
3. At the end of the course, what was your interest in the course content?
 - a. Very low
 - b. Low
 - c. Average
 - d. High
 - e. Very high
4. What level of effort did you put into the class?
 - a. Very low
 - b. Low
 - c. Average
 - d. High
 - e. Very high
5. What factor related to the instructor most influenced your learning in the course?
 - a. Instructor's manner of presentation
 - b. Instructor's teaching methods
 - c. Materials selected for the course
 - d. The instructor's method of grading
 - e. The instructor's personality

6. What factor related to you as a student most influenced your learning in the course?
 - a. My overall academic ability
 - b. The course's relationship to my career goals
 - c. My interest in the subject
 - d. The grade I will receive
 - e. The fact that the course was required
7. The instructor's teaching was...
 - a. Poor
 - b. Below average
 - c. Average
 - d. Above average
 - e. Excellent
8. My overall impression of this course was...
 - a. Poor
 - b. Below average
 - c. Average
 - d. Above average
 - e. Excellent
9. How much did you learn from this course?
 - a. Very little
 - b. A little
 - c. Some
 - d. Much
 - e. Very much
10. In this class, standard for student performance were...
 - a. Very low
 - b. Low
 - c. Average
 - d. High
 - e. Very high

11. The instructor was well-prepared for class.
- a. Strongly disagree
 - b. Disagree
 - c. Undecided
 - d. Agree
 - e. Strongly agree
12. Course content was related to general knowledge and experience external to the course.
- a. Strongly disagree
 - b. Disagree
 - c. Undecided
 - d. Agree
 - e. Strongly agree
13. Judging by presentations and answers to questions, the instructor displayed a clear understanding of the course topics.
- a. Strongly disagree
 - b. Disagree
 - c. Undecided
 - d. Agree
 - e. Strongly agree
14. Difficult concepts were explained in a helpful manner.
- a. Strongly disagree
 - b. Disagree
 - c. Undecided
 - d. Agree
 - e. Strongly agree
15. The instructor found alternative ways of explaining material when students didn't understand.
- a. Strongly disagree
 - b. Disagree
 - c. Undecided

- d. Agree
 - e. Strongly agree
16. Class presentations were intellectually stimulating.
- a. Strongly disagree
 - b. Disagree
 - c. Undecided
 - d. Agree
 - e. Strongly agree
17. The instructor caused me to think critically.
- a. Strongly disagree
 - b. Disagree
 - c. Undecided
 - d. Agree
 - e. Strongly agree
18. Course goals were clear.
- a. Strongly disagree
 - b. Disagree
 - c. Undecided
 - d. Agree
 - e. Strongly agree
19. There was agreement between announced goals of the course and what was actually taught.
- a. Strongly disagree
 - b. Disagree
 - c. Undecided
 - d. Agree
 - e. Strongly agree
20. Homework assignments and projects covered materials that had been presented.
- a. Strongly disagree
 - b. Disagree
 - c. Undecided

- d. Agree
 - e. Strongly agree
21. The grading system was clearly explained.
- a. Strongly disagree
 - b. Disagree
 - c. Undecided
 - d. Agree
 - e. Strongly agree
22. Grading in the course was based on how well students performed on assigned work.
- a. Strongly disagree
 - b. Disagree
 - c. Undecided
 - d. Agree
 - e. Strongly agree
23. Textbooks and other materials fit the goals of the course.
- a. Strongly disagree
 - b. Disagree
 - c. Undecided
 - d. Agree
 - e. Strongly agree
24. Textbooks and other materials helped in learning the course content.
- a. Strongly disagree
 - b. Disagree
 - c. Undecided
 - d. Agree
 - e. Strongly agree
25. The overall organization of the course (relationship among lectures, readings, and classroom activities) contributed to learning.
- a. Strongly disagree
 - b. Disagree

- c. Undecided
- d. Agree
- e. Strongly agree

Appendix B

Parallel Analysis Results

Factor	Actual data	Randomly generated data	
	Initial eigenvalues	Mean eigenvalues	95 th percentile eigenvalues
1	12.249	1.107	1.126
2	1.153	1.089	1.102
3	1.087	1.075	1.086
4	.682	1.062	1.073
5	.567	1.050	1.060
6	.415	1.039	1.048
7	.385	1.029	1.038
8	.342	1.019	1.027
9	.323	1.009	1.018
10	.290	.999	1.007
11	.246	.989	.998
12	.219	.980	.988
13	.203	.970	.978
14	.189	.960	.968
15	.150	.949	.959
16	.138	.938	.947
17	.131	.926	.937
18	.128	.913	.925
19	.104	.896	.911

Appendix C

Principal Component Factor Analysis Scree Plot

Scree Plot

Appendix D

Correlations between predictor and outcome variables

Variable	1	2	3	4	5	6	7	8	9
1. Electivity	-								
2. Prior interest	-.037*	-							
3. Effort	.050**	.276**	-						
4. Class Level	-.127**	-.005	.001	-					
5. Class time start	-.029	-.018	.010	.515**	-				
6. Class length	-.045**	-.025	.040**	.404**	.322**	-			
7. Class enrollment	.079**	-.071**	-.111**	-.366**	-.359**	-.399**	-		
8. Class meeting pattern	.010	.037*	-.037*	-.487**	-.513**	-.823**	.487**	-	
9. Student gender	.014	.027	.039*	.038*	.027	.113**	-.147**	-.152**	-
10. Student race	-.009	-.049**	-.023	.002	.036*	-.009	.010	.022	-.002
11. Student grade	-.004	.056**	.075**	.146**	.114*	.183**	-.161**	-.223**	.097**
12. Student age	-.069**	.054**	.104**	.330**	.382**	.285**	-.220**	-.338**	-.078**
13. Instructor race	.034*	.013	-.081**	.032*	.004	-.145**	.021	.146**	-.036*
14. Instructor position	-.008	-.012	.095**	.282**	.206**	.306**	-.245**	-.379**	.052**
15. Instructor gender	.023	.064**	.049**	-.175**	-.219**	.052**	-.120**	-.033*	.132**
16. Total rating	-.013	.176**	.350**	-.031*	-.052**	-.023	-.030*	.059**	.005
17. Factor one	< .001	.126**	.269**	-.043**	-.060**	-.035*	-.007	.070**	.001
18. Factor two	-.031*	.227**	.436**	-.014	-.041**	.003	-.061**	.031*	-.002
19. Factor three	-.011	.155**	.265**	-.002	-.012**	-.031*	-.033*	.054**	.042**

Appendix D continued

Correlations between predictor and outcome variables

Variable	10	11	12	13	14	15	16	17	18
10. Student race	-								
11. Student grade	-.113**	-							
12. Student age	.062**	.058**	-						
13. Instructor race	.040*	-.048**	-.027	-					
14. Instructor position	-.010	.056**	.150**	-.070**	-				
15. Instructor gender	-.055**	.080**	-.168**	-.109**	.063**	-			
16. Total rating	.015	.088**	.039*	-.092**	.096**	-.020	-		
17. Factor one	.008	.087**	.010	-.092**	.086**	-.023	.973**	-	
18. Factor two	.019	.085**	.072**	-.099**	.114**	-.020	.933**	.845**	-
19. Factor three	.024	.053**	.052**	-.022	.034*	.004	.712**	.634**	.587**

Appendix E

Overall Course Rating: Unconditional Model

Fixed Effect	Coefficient	S.E.	<i>t</i> -ratio	<i>d.f.</i>	<i>p</i> -value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	62.269825	0.613754	101.457	177	<0.001

Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	4.42861	19.61256	325	685.33296	<0.001
level-1, e	11.42503	130.53128			
INTRCPT1/INTRCPT2, u_{00}	6.79086	46.11575	177	792.62596	<0.001

Appendix F

Overall Course Rating: Student-level Model

Fixed Effect	Coefficient	S.E.	t-ratio	d.f.	p-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	62.046706	1.176772	52.726	149	<0.001
For Q1 slope, π_1					
For INTRCPT2, β_{10}					
INTRCPT3, γ_{100}	0.039442	0.998385	0.040	94	0.969
For Q2 slope, π_2					
For INTRCPT2, β_{20}					
INTRCPT3, γ_{200}	1.188446	0.218008	5.451	415	<0.001
For Q4 slope, π_3					
For INTRCPT2, β_{30}					
INTRCPT3, γ_{300}	5.130092	0.249225	20.584	2269	<0.001
For STUD_GEN slope, π_4					
For INTRCPT2, β_{40}					
INTRCPT3, γ_{400}	-0.229298	0.496449	-0.462	177	0.645
For STUD_RAC slope, π_5					
For INTRCPT2, β_{50}					
INTRCPT3, γ_{500}	0.949452	0.573539	1.655	177	0.100
For STUD_AGE slope, π_6					
For INTRCPT2, β_{60}					
INTRCPT3, γ_{600}	0.047432	0.026598	1.783	2200	0.075
Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	2.34276	5.48850	90	129.45462	0.004
STUD_GEN slope,r4	2.82531	7.98239	90	122.91593	0.012
STUD_RAC slope,r5	1.76978	3.13211	90	130.09451	0.004
level-1, e	10.43263	108.83983			
INTRCPT1/INTRCPT2,u00	7.19154	51.71823	135	545.86836	<0.001
STUD_GEN/INTRCPT2,u40	1.87265	3.50681	135	154.40964	0.121
STUD_RAC/INTRCPT2,u50	2.37382	5.63503	135	143.56662	0.291

Appendix G

Overall Course Rating: Course-level Model

Fixed Effect	Coefficient	S.E.	t-ratio	d.f.	p-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	61.624101	1.459144	42.233	177	<0.001
For COUR_ENR, β_{01}					
INTRCPT3, γ_{010}	-0.029891	0.036649	-0.816	470	0.415
For CO_LEVEL, β_{02}					
INTRCPT3, γ_{020}	-0.235790	1.006096	-0.234	470	0.815
For CO LENGT, β_{03}					
INTRCPT3, γ_{030}	0.011000	0.014815	0.742	241	0.459
For CO_NUMBE, β_{04}					
INTRCPT3, γ_{040}	2.345651	1.755630	1.336	470	0.182
For CO_START, β_{05}					
INTRCPT3, γ_{050}	-0.196048	0.311453	-0.629	470	0.529
For Q1 slope, π_1					
For INTRCPT2, β_{10}					
INTRCPT3, γ_{100}	0.098320	0.984820	0.100	110	0.921
For Q2 slope, π_2					
For INTRCPT2, β_{20}					
INTRCPT3, γ_{200}	1.171649	0.216055	5.423	544	<0.001
For Q4 slope, π_3					
For INTRCPT2, β_{30}					
INTRCPT3, γ_{300}	5.116206	0.247802	20.646	2269	<0.001
For STUD_GEN slope, π_4					
For INTRCPT2, β_{40}					
INTRCPT3, γ_{400}	-0.217300	0.497466	-0.437	177	0.663
For STUD_RAC slope, π_5					
For INTRCPT2, β_{50}					
INTRCPT3, γ_{500}	0.936471	0.572775	1.635	177	0.104
For STUD_AGE slope, π_6					
For INTRCPT2, β_{60}					
INTRCPT3, γ_{600}	0.055083	0.027439	2.007	1326	0.045
Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	2.38249	5.67628	85	130.60936	0.001
STUD_GEN slope,r4	2.81042	7.89843	90	122.80605	0.012
STUD_RAC slope,r5	1.75816	3.09113	90	129.93481	0.004
level-1, e	10.43597	108.90956			
INTRCPT1/INTRCPT2,u00	7.12562	50.77441	135	535.11694	<0.001

STUD_GEN/INTRCPT2,u40	1.82016	3.31299	135	153.61449	0.131
STUD_RAC/INTRCPT2,u50	2.38031	5.66589	135	143.51834	0.291

Appendix H

Overall Course Rating: Instructor-level Model

Fixed Effect	Coefficient	S.E.	<i>t</i> -ratio	<i>d.f.</i>	<i>p</i> -value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	63.147891	1.656400	38.124	174	<0.001
INST_RAC, γ_{001}	-3.352417	1.726635	-1.942	174	0.054
INST_GEN, γ_{002}	-2.003941	1.350479	-1.484	174	0.140
INST_POS, γ_{003}	-0.386522	0.553527	-0.698	174	0.486
For COUR_ENR, β_{01}					
INTRCPT3, γ_{010}	-0.032252	0.036671	-0.880	470	0.380
For CO_LEVEL, β_{02}					
INTRCPT3, γ_{020}	-0.107249	1.013172	-0.106	470	0.916
For CO LENGT, β_{03}					
INTRCPT3, γ_{030}	0.010499	0.014889	0.705	221	0.481
For CO_NUMBE, β_{04}					
INTRCPT3, γ_{040}	2.492160	1.779240	1.401	470	0.162
For CO_START, β_{05}					
INTRCPT3, γ_{050}	-0.214616	0.306345	-0.701	470	0.484
For Q1 slope, π_1					
For INTRCPT2, β_{10}					
INTRCPT3, γ_{100}	0.145922	0.982332	0.149	114	0.882
For Q2 slope, π_2					
For INTRCPT2, β_{20}					
INTRCPT3, γ_{200}	1.158637	0.217158	5.335	496	<0.001
For Q4 slope, π_3					
For INTRCPT2, β_{30}					
INTRCPT3, γ_{300}	5.104322	0.247550	20.619	2269	<0.001
For STUD_GEN slope, π_4					
For INTRCPT2, β_{40}					
INTRCPT3, γ_{400}	-1.146434	0.788257	-1.454	176	0.148
INST_GEN, γ_{401}	1.525687	0.999114	1.527	176	0.129
For STUD_RAC slope, π_5					
For INTRCPT2, β_{50}					
INTRCPT3, γ_{500}	0.695914	0.623645	1.116	176	0.266
INST_RAC, γ_{501}	1.588481	1.496989	1.061	176	0.290
For STUD_AGE slope, π_6					
For INTRCPT2, β_{60}					
INTRCPT3, γ_{600}	0.055100	0.027323	2.017	1702	0.044

Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	2.39401	5.73129	85	130.80167	0.001
STUD_GEN slope,r4	2.85921	8.17506	90	122.96002	0.012
STUD_RAC slope,r5	1.77955	3.16678	90	129.97591	0.004
level-1, e	10.43183	108.82305			
INTRCPT1/INTRCPT2,u00	6.98932	48.85055	132	511.76131	<0.001
STUD_GEN/INTRCPT2,u40	1.62741	2.64847	134	149.65540	0.168
STUD_RAC/INTRCPT2,u50	2.31197	5.34521	134	141.53874	0.311

Appendix I

Factor One: Unconditional Model

Fixed Effect	Coefficient	S.E.	<i>t</i> -ratio	<i>d.f.</i>	<i>p</i> -value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	37.072404	0.354759	104.500	177	<0.001
Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	2.49648	6.23244	325	652.63261	<0.001
level-1, e	6.73081	45.30380			
INTRCPT1/INTRCPT2,u00	3.96111	15.69036	177	811.96319	<0.001

Appendix J

Factor One: Student-level Model

Fixed Effect	Coefficient	S.E.	t-ratio	d.f.	p-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	36.917943	0.658245	56.085	177	<0.001
For Q1 slope, π_1					
For INTRCPT2, β_{10}					
INTRCPT3, γ_{100}	0.131892	0.569183	0.232	184	0.817
For Q2 slope, π_2					
For INTRCPT2, β_{20}					
INTRCPT3, γ_{200}	0.461733	0.135584	3.406	258	<0.001
For Q4 slope, π_3					
For INTRCPT2, β_{30}					
INTRCPT3, γ_{300}	2.365533	0.151840	15.579	2269	<0.001
For STUD_GEN slope, π_4					
For INTRCPT2, β_{40}					
INTRCPT3, γ_{400}	-0.165010	0.294734	-0.560	177	0.576
For STUD_RAC slope, π_5					
For INTRCPT2, β_{50}					
INTRCPT3, γ_{500}	0.383928	0.341183	1.125	177	0.262
For STUD_AGE slope, π_6					
For INTRCPT2, β_{60}					
INTRCPT3, γ_{600}	0.016860	0.016267	1.037	1413	0.300
Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	1.19105	1.41861	90	116.45507	0.032
STUD_GEN slope,r4	1.79041	3.20555	90	122.82129	0.012
STUD_RAC slope,r5	1.01247	1.02509	90	126.23410	0.007
level-1, e	6.35405	40.37393			
INTRCPT1/INTRCPT2,u00	4.33066	18.75459	135	556.18579	<0.001
STUD_GEN/INTRCPT2,u40	0.95270	0.90764	135	142.12550	0.320
STUD_RAC/INTRCPT2,u50	1.32320	1.75085	135	140.96460	0.345

Appendix K

Factor One: Course-level Model

Fixed Effect	Coefficient	S.E.	t-ratio	d.f.	p-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	36.652323	0.811575	45.162	177	<0.001
For COUR_ENR, β_{01}					
INTRCPT3, γ_{010}	-0.007574	0.021733	-0.349	470	0.728
For CO_LEVEL, β_{02}					
INTRCPT3, γ_{020}	-0.115170	0.587778	-0.196	470	0.845
For CO LENGT, β_{03}					
INTRCPT3, γ_{030}	0.005870	0.008755	0.670	288	0.503
For CO_NUMBE, β_{04}					
INTRCPT3, γ_{040}	1.329806	1.054112	1.262	470	0.208
For CO_START, β_{05}					
INTRCPT3, γ_{050}	-0.095775	0.183776	-0.521	470	0.603
For Q1 slope, π_1					
For INTRCPT2, β_{10}					
INTRCPT3, γ_{100}	0.153121	0.561717	0.273	237	0.785
For Q2 slope, π_2					
For INTRCPT2, β_{20}					
INTRCPT3, γ_{200}	0.452861	0.134413	3.369	313	<0.001
For Q4 slope, π_3					
For INTRCPT2, β_{30}					
INTRCPT3, γ_{300}	2.359128	0.151089	15.614	2269	<0.001
For STUD_GEN slope, π_4					
For INTRCPT2, β_{40}					
INTRCPT3, γ_{400}	-0.149174	0.295194	-0.505	177	0.614
For STUD_RAC slope, π_5					
For INTRCPT2, β_{50}					
INTRCPT3, γ_{500}	0.378154	0.340566	1.110	177	0.268
For STUD_AGE slope, π_6					
For INTRCPT2, β_{60}					
INTRCPT3, γ_{600}	0.021815	0.016859	1.294	806	0.196
Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	1.12094	1.25651	85	116.93933	0.012
STUD_GEN slope,r4	1.82919	3.34594	90	122.97465	0.012
STUD_RAC slope,r5	1.01669	1.03367	90	126.07219	0.007
level-1, e	6.35663	40.40672			
INTRCPT1/INTRCPT2, u_{00}	4.31231	18.59606	135	556.55811	<0.001

STUD_GEN/INTRCPT2, u_{40}	0.89643	0.80359	135	140.61590	0.353
STUD_RAC/INTRCPT2, u_{50}	1.31803	1.73721	135	140.79293	0.349

Appendix L

Factor One: Instructor-level Model

Fixed Effect	Coefficient	S.E.	t-ratio	d.f.	p-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	37.517746	0.960051	39.079	174	<0.001
INST_RAC, γ_{001}	-2.033708	1.034509	-1.966	174	0.051
INST_GEN, γ_{002}	-1.101753	0.813391	-1.355	174	0.177
INST_POS, γ_{003}	-0.283224	0.330437	-0.857	174	0.393
For COUR_ENR, β_{01}					
INTRCPT3, γ_{010}	-0.008926	0.021869	-0.408	470	0.683
For CO_LEVEL, β_{02}					
INTRCPT3, γ_{020}	-0.024582	0.595886	-0.041	470	0.967
For CO LENGT, β_{03}					
INTRCPT3, γ_{030}	0.005626	0.008808	0.639	268	0.524
For CO_NUMBE, β_{04}					
INTRCPT3, γ_{040}	1.373491	1.072720	1.280	470	0.201
For CO_START, β_{05}					
INTRCPT3, γ_{050}	-0.107851	0.181797	-0.593	470	0.553
For Q1 slope, π_1					
For INTRCPT2, β_{10}					
INTRCPT3, γ_{100}	0.180736	0.561038	0.322	244	0.748
For Q2 slope, π_2					
For INTRCPT2, β_{20}					
INTRCPT3, γ_{200}	0.445997	0.134851	3.307	305	0.001
For Q4 slope, π_3					
For INTRCPT2, β_{30}					
INTRCPT3, γ_{300}	2.353557	0.150889	15.598	2269	<0.001
For STUD_GEN slope, π_4					
For INTRCPT2, β_{40}					
INTRCPT3, γ_{400}	-0.596814	0.471532	-1.266	176	0.207
INST_GEN, γ_{401}	0.744849	0.599123	1.243	176	0.215
For STUD_RAC slope, π_5					
For INTRCPT2, β_{50}					
INTRCPT3, γ_{500}	0.214498	0.375178	0.572	176	0.568
INST_RAC, γ_{501}	1.068865	0.906544	1.179	176	0.240
For STUD_AGE slope, π_6					
For INTRCPT2, β_{60}					
INTRCPT3, γ_{600}	0.021574	0.016783	1.285	995	0.199

Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	1.18055	1.39371	85	117.21178	0.012
STUD_GEN slope,r4	1.79240	3.21270	90	122.85968	0.012
STUD_RAC slope,r5	0.96784	0.93671	90	126.09242	0.007
level-1, e	6.35458	40.38071			
INTRCPT1/INTRCPT2,u00	4.22725	17.86968	132	524.04003	<0.001
STUD_GEN/INTRCPT2,u40	0.80916	0.65474	134	138.47503	0.378
STUD_RAC/INTRCPT2,u50	1.29851	1.68612	134	139.18309	0.362

Appendix M

Factor Two: Unconditional Model

Fixed Effect	Coefficient	S.E.	<i>t</i> -ratio	<i>d.f.</i>	<i>p</i> -value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	18.630439	0.224352	83.041	177	<0.001
Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	1.58231	2.50369	325	663.07635	<0.001
level-1, e	4.20570	17.68788			
INTRCPT1/INTRCPT2,u00	2.44499	5.97800	177	777.35771	<0.001

Appendix N

Factor Two: Student-level Model

Fixed Effect	Coefficient	S.E.	t-ratio	d.f.	p-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	18.715126	0.432720	43.250	94	<0.001
For Q1 slope, π_1					
For INTRCPT2, β_{10}					
INTRCPT3, γ_{100}	-0.099284	0.347726	-0.286	100	0.776
For Q2 slope, π_2					
For INTRCPT2, β_{20}					
INTRCPT3, γ_{200}	0.551422	0.072584	7.597	2269	<0.001
For Q4 slope, π_3					
For INTRCPT2, β_{30}					
INTRCPT3, γ_{300}	2.324281	0.087991	26.415	2269	<0.001
For STUD_GEN slope, π_4					
For INTRCPT2, β_{40}					
INTRCPT3, γ_{400}	-0.197861	0.191608	-1.033	177	0.303
For STUD_RAC slope, π_5					
For INTRCPT2, β_{50}					
INTRCPT3, γ_{500}	0.386252	0.198953	1.941	177	0.054
For STUD_AGE slope, π_6					
For INTRCPT2, β_{60}					
INTRCPT3, γ_{600}	0.026033	0.009272	2.808	2269	0.005
Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	0.98232	0.96496	90	126.34332	0.007
STUD_GEN slope,r4	0.82908	0.68738	90	105.98717	0.120
STUD_RAC slope,r5	0.66812	0.44638	90	124.60177	0.009
level-1, e	3.66878	13.45993			
INTRCPT1/INTRCPT2,u00	2.51342	6.31731	135	502.63906	<0.001
STUD_GEN/INTRCPT2,u40	1.00135	1.00269	135	177.04810	0.009
STUD_RAC/INTRCPT2,u50	0.86358	0.74578	135	144.23487	0.277

Appendix O

Factor Two: Course-level Model

Fixed Effect	Coefficient	S.E.	t-ratio	d.f.	p-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	18.628428	0.518757	35.910	150	<0.001
For COUR_ENR, β_{01}					
INTRCPT3, γ_{010}	-0.024030	0.012505	-1.922	470	0.055
For CO_LEVEL, β_{02}					
INTRCPT3, γ_{020}	-0.115179	0.344137	-0.335	470	0.738
For CO_LENGT, β_{03}					
INTRCPT3, γ_{030}	0.003896	0.005322	0.732	142	0.465
For CO_NUMBE, β_{04}					
INTRCPT3, γ_{040}	0.725671	0.607901	1.194	470	0.233
For CO_START, β_{05}					
INTRCPT3, γ_{050}	-0.108423	0.108357	-1.001	470	0.318
For Q1 slope, π_1					
For INTRCPT2, β_{10}					
INTRCPT3, γ_{100}	-0.060603	0.343669	-0.176	115	0.860
For Q2 slope, π_2					
For INTRCPT2, β_{20}					
INTRCPT3, γ_{200}	0.542796	0.072343	7.503	2269	<0.001
For Q4 slope, π_3					
For INTRCPT2, β_{30}					
INTRCPT3, γ_{300}	2.316241	0.087330	26.523	2269	<0.001
For STUD_GEN slope, π_4					
For INTRCPT2, β_{40}					
INTRCPT3, γ_{400}	-0.206341	0.191807	-1.076	177	0.283
For STUD_RAC slope, π_5					
For INTRCPT2, β_{50}					
INTRCPT3, γ_{500}	0.384673	0.198479	1.938	177	0.054
For STUD_AGE slope, π_6					
For INTRCPT2, β_{60}					
INTRCPT3, γ_{600}	0.028148	0.009576	2.939	1688	0.003
Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	0.96363	0.92858	85	125.99908	0.003
STUD_GEN slope,r4	0.81384	0.66234	90	105.77197	0.123
STUD_RAC slope,r5	0.60584	0.36704	90	124.09834	0.010
level-1, e	3.67100	13.47627			
INTRCPT1/INTRCPT2,u00	2.49110	6.20560	135	501.20215	<0.001

STUD_GEN/INTRCPT2,u40	0.98811	0.97636	135	176.64555	0.009
STUD_RAC/INTRCPT2,u50	0.86717	0.75198	135	145.14623	0.260

Appendix P

Factor Two: Instructor-level Model

Fixed Effect	Coefficient	S.E.	t-ratio	d.f.	p-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	19.288537	0.588549	32.773	174	<0.001
INST_RAC, γ_{001}	-1.178311	0.601049	-1.960	174	0.052
INST_GEN, γ_{002}	-0.884276	0.472238	-1.873	174	0.063
INST_POS, γ_{003}	-0.060463	0.190851	-0.317	174	0.752
For COUR_ENR, β_{01}					
INTRCPT3, γ_{010}	-0.025138	0.012505	-2.010	470	0.045
For CO_LEVEL, β_{02}					
INTRCPT3, γ_{020}	-0.103145	0.348262	-0.296	470	0.767
For CO LENGT, β_{03}					
INTRCPT3, γ_{030}	0.003633	0.005306	0.685	145	0.495
For CO_NUMBE, β_{04}					
INTRCPT3, γ_{040}	0.795639	0.616326	1.291	470	0.197
For CO_START, β_{05}					
INTRCPT3, γ_{050}	-0.117751	0.106450	-1.106	470	0.269
For Q1 slope, π_1					
For INTRCPT2, β_{10}					
INTRCPT3, γ_{100}	-0.044601	0.340456	-0.131	127	0.896
For Q2 slope, π_2					
For INTRCPT2, β_{20}					
INTRCPT3, γ_{200}	0.539963	0.072661	7.431	2269	<0.001
For Q4 slope, π_3					
For INTRCPT2, β_{30}					
INTRCPT3, γ_{300}	2.312033	0.087124	26.537	2269	<0.001
For STUD_GEN slope, π_4					
For INTRCPT2, β_{40}					
INTRCPT3, γ_{400}	-0.580311	0.300214	-1.933	176	0.055
INST_GEN, γ_{401}	0.618120	0.372808	1.658	176	0.099
For STUD_RAC slope, π_5					
For INTRCPT2, β_{50}					
INTRCPT3, γ_{500}	0.317657	0.215831	1.472	176	0.143
INST_RAC, γ_{501}	0.446719	0.527194	0.847	176	0.398
For STUD_AGE slope, π_6					
For INTRCPT2, β_{60}					
INTRCPT3, γ_{600}	0.028017	0.009534	2.939	2265	0.003

Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	0.98155	0.96344	85	126.11534	0.003
STUD_GEN slope,r4	0.87260	0.76142	90	106.01138	0.120
STUD_RAC slope,r5	0.57796	0.33404	90	123.95722	0.010
level-1, e	3.66993	13.46840			
INTRCPT1/INTRCPT2,u00	2.43871	5.94732	132	478.14507	<0.001
STUD_GEN/INTRCPT2,u40	0.90364	0.81657	134	171.78082	0.015
STUD_RAC/INTRCPT2,u50	0.87719	0.76946	134	144.58631	0.251

Appendix Q

Factor Three: Unconditional Model

Fixed Effect	Coefficient	S.E.	<i>t</i> -ratio	<i>d.f.</i>	<i>p</i> -value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	6.557268	0.069357	94.543	177	<0.001
Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	0.55147	0.30412	325	581.14000	<0.001
level-1, e	1.60442	2.57417			
INTRCPT1/INTRCPT2,u00	0.66588	0.44340	177	485.83005	<0.001

Appendix R

Factor Three: Student-level Model

Fixed Effect	Coefficient	S.E.	t-ratio	d.f.	p-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	6.437290	0.168138	38.286	69	<0.001
For Q1 slope, π_1					
For INTRCPT2, β_{10}					
INTRCPT3, γ_{100}	-0.004213	0.153399	-0.027	53	0.978
For Q2 slope, π_2					
For INTRCPT2, β_{20}					
INTRCPT3, γ_{200}	0.173964	0.029612	5.875	2269	<0.001
For Q4 slope, π_3					
For INTRCPT2, β_{30}					
INTRCPT3, γ_{300}	0.449188	0.035654	12.598	2269	<0.001
For STUD_GEN slope, π_4					
For INTRCPT2, β_{40}					
INTRCPT3, γ_{400}	0.114858	0.068176	1.685	177	0.094
For STUD_RAC slope, π_5					
For INTRCPT2, β_{50}					
INTRCPT3, γ_{500}	0.175340	0.079160	2.215	177	0.028
For STUD_AGE slope, π_6					
For INTRCPT2, β_{60}					
INTRCPT3, γ_{600}	0.005165	0.003659	1.412	2269	0.158
Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	0.45681	0.20867	90	168.97624	<0.001
STUD_GEN slope,r4	0.41076	0.16873	90	159.39733	<0.001
STUD_RAC slope,r5	0.19477	0.03794	90	148.57599	<0.001
level-1, e	1.53875	2.36776			
INTRCPT1/INTRCPT2,u00	0.67655	0.45772	135	282.89794	<0.001
STUD_GEN/INTRCPT2,u40	0.09912	0.00983	135	133.61828	>.500
STUD_RAC/INTRCPT2,u50	0.21468	0.04609	135	125.54387	>.500

Appendix S

Factor Three: Course-level Model

Fixed Effect	Coefficient	S.E.	t-ratio	d.f.	p-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	6.335820	0.217360	29.149	55	<0.001
For COUR_ENR, β_{01}					
INTRCPT3, γ_{010}	-0.000806	0.004935	-0.163	470	0.870
For CO_LEVEL, β_{02}					
INTRCPT3, γ_{020}	-0.015000	0.129328	-0.116	470	0.908
For CO LENGT, β_{03}					
INTRCPT3, γ_{030}	0.001571	0.001785	0.880	470	0.379
For CO_NUMBE, β_{04}					
INTRCPT3, γ_{040}	0.424311	0.224973	1.886	470	0.060
For CO_START, β_{05}					
INTRCPT3, γ_{050}	0.022329	0.037629	0.593	470	0.553
For Q1 slope, π_1					
For INTRCPT2, β_{10}					
INTRCPT3, γ_{100}	-0.001667	0.153242	-0.011	55	0.991
For Q2 slope, π_2					
For INTRCPT2, β_{20}					
INTRCPT3, γ_{200}	0.171507	0.029595	5.795	2269	<0.001
For Q4 slope, π_3					
For INTRCPT2, β_{30}					
INTRCPT3, γ_{300}	0.449655	0.035694	12.597	2269	<0.001
For STUD_GEN slope, π_4					
For INTRCPT2, β_{40}					
INTRCPT3, γ_{400}	0.128361	0.068439	1.876	177	0.062
For STUD_RAC slope, π_5					
For INTRCPT2, β_{50}					
INTRCPT3, γ_{500}	0.171535	0.079010	2.171	177	0.031
For STUD_AGE slope, π_6					
For INTRCPT2, β_{60}					
INTRCPT3, γ_{600}	0.006431	0.003787	1.698	2269	0.090
Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	0.45715	0.20899	85	170.19997	<0.001
STUD_GEN slope,r4	0.41358	0.17105	90	159.53919	<0.001
STUD_RAC slope,r5	0.18006	0.03242	90	148.71880	<0.001
level-1, e	1.53858	2.36722			

INTRCPT1/INTRCPT2, u_{00}	0.66483	0.44201	135	277.70420	<0.001
STUD_GEN/INTRCPT2, u_{40}	0.09694	0.00940	135	133.49194	>.500
STUD_RAC/INTRCPT2, u_{50}	0.23924	0.05724	135	126.09478	>.500

Appendix T

Factor Three: Instructor-level Model

Fixed Effect	Coefficient	S.E.	t-ratio	d.f.	p-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	6.345531	0.218273	29.072	174	<0.001
INST_RAC, γ_{001}	-0.132301	0.193781	-0.683	174	0.496
INST_GEN, γ_{002}	-0.038274	0.169240	-0.226	174	0.821
INST_POS, γ_{003}	-0.026378	0.062681	-0.421	174	0.674
For COUR_ENR, β_{01}					
INTRCPT3, γ_{010}	-0.000399	0.004913	-0.081	470	0.935
For CO_LEVEL, β_{02}					
INTRCPT3, γ_{020}	0.015607	0.128962	0.121	470	0.904
For CO LENGT, β_{03}					
INTRCPT3, γ_{030}	0.001610	0.001805	0.892	470	0.373
For CO_NUMBE, β_{04}					
INTRCPT3, γ_{040}	0.450469	0.228769	1.969	470	0.050
For CO_START, β_{05}					
INTRCPT3, γ_{050}	0.027355	0.037855	0.723	470	0.470
For Q1 slope, π_1					
For INTRCPT2, β_{10}					
INTRCPT3, γ_{100}	0.002391	0.155168	0.015	52	0.988
For Q2 slope, π_2					
For INTRCPT2, β_{20}					
INTRCPT3, γ_{200}	0.169182	0.029766	5.684	2269	<0.001
For Q4 slope, π_3					
For INTRCPT2, β_{30}					
INTRCPT3, γ_{300}	0.448847	0.035712	12.568	2269	<0.001
For STUD_GEN slope, π_4					
For INTRCPT2, β_{40}					
INTRCPT3, γ_{400}	0.013922	0.116432	0.120	176	0.905
INST_GEN, γ_{401}	0.179486	0.152418	1.178	176	0.241
For STUD_RAC slope, π_5					
For INTRCPT2, β_{50}					
INTRCPT3, γ_{500}	0.167729	0.084960	1.974	176	0.050
INST_RAC, γ_{501}	0.038978	0.210180	0.185	176	0.853
For STUD_AGE slope, π_6					
For INTRCPT2, β_{60}					
INTRCPT3, γ_{600}	0.006682	0.003794	1.761	2269	0.078

Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	0.45497	0.20700	85	170.32940	<0.001
STUD_GEN slope,r4	0.39260	0.15414	90	159.48579	<0.001
STUD_RAC slope,r5	0.18124	0.03285	90	148.75638	<0.001
level-1, e	1.53841	2.36670			
INTRCPT1/INTRCPT2,u00	0.65884	0.43407	132	275.11459	<0.001
STUD_GEN/INTRCPT2,u40	0.08735	0.00763	134	131.84800	>.500
STUD_RAC/INTRCPT2,u50	0.24304	0.05907	134	126.10298	>.500

Appendix U

Student Grade as an Outcome: Unconditional Model

Fixed Effect	Coefficient	S.E.	<i>t</i> -ratio	<i>d.f.</i>	<i>p</i> -value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	3.766542	0.024292	155.054	54	<0.001
Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	0.15905	0.02530	325	639.29488	<0.001
level-1, e	0.39898	0.15918			
INTRCPT1/INTRCPT2,u00	0.19560	0.03826	177	568.63247	<0.001

Appendix V

Student Grade as an Outcome: Control Variables Model

Fixed Effect	Coefficient	S.E.	t-ratio	d.f.	p-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	3.768108	0.054446	69.208	174	<0.001
INST_RAC, γ_{001}	0.011906	0.047249	0.252	174	0.801
INST_GEN, γ_{002}	0.024381	0.035457	0.688	174	0.493
INST_POS, γ_{003}	-0.046434	0.017603	-2.638	174	0.009
For COUR_ENR, β_{01}					
INTRCPT3, γ_{010}	-0.002697	0.001246	-2.165	292	0.031
For CO_LEVEL, β_{02}					
INTRCPT3, γ_{020}	0.080339	0.033331	2.410	292	0.017
For CO LENGT, β_{03}					
INTRCPT3, γ_{030}	-0.000015	0.000483	-0.031	292	0.975
For CO_NUMBE, β_{04}					
INTRCPT3, γ_{040}	-0.202942	0.058726	-3.456	292	<0.001
For CO_START, β_{05}					
INTRCPT3, γ_{050}	0.003530	0.014178	0.249	32	0.805
For Q1 slope, π_1					
For INTRCPT2, β_{10}					
INTRCPT3, γ_{100}	-0.006570	0.031549	-0.208	2598	0.835
For Q2 slope, π_2					
For INTRCPT2, β_{20}					
INTRCPT3, γ_{200}	0.012395	0.008341	1.486	288	0.138
For Q4 slope, π_3					
For INTRCPT2, β_{30}					
INTRCPT3, γ_{300}	0.054915	0.009762	5.625	789	<0.001
For STUD_GEN slope, π_4					
For INTRCPT2, β_{40}					
INTRCPT3, γ_{400}	0.033807	0.016855	2.006	3573	0.045
For STUD_RAC slope, π_5					
For INTRCPT2, β_{50}					
INTRCPT3, γ_{500}	-0.136709	0.018821	-7.264	3573	<0.001
For STUD_AGE slope, π_6					
For INTRCPT2, β_{60}					
INTRCPT3, γ_{600}	-0.000918	0.001136	-0.808	136	0.420
Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	0.14845	0.02204	320	606.54638	<0.001
level-1, e	0.39232	0.15391			

INTRCPT1/INTRCPT2,u00	0.16031	0.02570	174	484.41495	<0.001
-----------------------	---------	---------	-----	-----------	--------

Appendix W

Student Grade as an Outcome: Student Rating Predictor Model

Fixed Effect	Coefficient	S.E.	t-ratio	d.f.	p-value
For INTRCPT1, π_0					
For INTRCPT2, β_{00}					
INTRCPT3, γ_{000}	3.766640	0.054182	69.519	174	<0.001
INST_RAC, γ_{001}	0.017861	0.046792	0.382	174	0.703
INST_GEN, γ_{002}	0.027043	0.035282	0.766	174	0.444
INST_POS, γ_{003}	-0.045804	0.017611	-2.601	174	0.010
For COUR_ENR, β_{01}					
INTRCPT3, γ_{010}	-0.002594	0.001239	-2.093	292	0.037
For CO_LEVEL, β_{02}					
INTRCPT3, γ_{020}	0.081028	0.033066	2.450	292	0.015
For CO LENGT, β_{03}					
INTRCPT3, γ_{030}	-0.000041	0.000479	-0.085	292	0.932
For CO_NUMBE, β_{04}					
INTRCPT3, γ_{040}	-0.208636	0.058555	-3.563	292	<0.001
For CO_START, β_{05}					
INTRCPT3, γ_{050}	0.003888	0.014277	0.272	31	0.787
For Q1 slope, π_1					
For INTRCPT2, β_{10}					
INTRCPT3, γ_{100}	-0.006812	0.031328	-0.217	3572	0.828
For Q2 slope, π_2					
For INTRCPT2, β_{20}					
INTRCPT3, γ_{200}	0.010044	0.008231	1.220	400	0.223
For Q4 slope, π_3					
For INTRCPT2, β_{30}					
INTRCPT3, γ_{300}	0.044352	0.010000	4.435	3199	<0.001
For STUD_GEN slope, π_4					
For INTRCPT2, β_{40}					
INTRCPT3, γ_{400}	0.034131	0.016850	2.026	3572	0.043
For STUD_RAC slope, π_5					
For INTRCPT2, β_{50}					
INTRCPT3, γ_{500}	-0.138404	0.018806	-7.360	3572	<0.001
For STUD_AGE slope, π_6					
For INTRCPT2, β_{60}					
INTRCPT3, γ_{600}	-0.001023	0.001139	-0.898	129	0.371
For TOTALEVA slope, π_7					
For INTRCPT2, β_{70}					
INTRCPT3, γ_{700}	0.002032	0.000634	3.202	1014	0.001

Random Effect	S.D.	Variance	d.f.	χ^2	p-value
INTRCPT1,r0	0.14812	0.02194	320	605.59988	<0.001
level-1, e	0.39186	0.15356			
INTRCPT1/INTRCPT2,u00	0.15787	0.02492	174	476.04149	<0.001

CURRICULUM VITAE
Timothy Michael Sauer

Home Address:
3907 Glen Oak Drive
Louisville, Kentucky 40218

phone: 502.387.0440
timothysauer@gmail.com

EDUCATION

Ph.D. University of Louisville. Educational Leadership & Organizational Development: Evaluation, May 2012.
Dissertation: Predictors of Student Course Evaluations.

B.A. Bellarmine University. Psychology & Criminal Justice, May 2006
summa cum laude

PROFESSIONAL EXPERIENCE

Independent Research and Evaluation Consultant **2009-present**
Louisville, Kentucky
 Provide support in quantitative research methodology, program evaluation and advanced statistics.

Graduate Research and Teaching Assistant **2010-2011**
College of Education and Human Development
University of Louisville

Worked as the teaching assistant for graduate level courses in structural equation modeling and hierarchical linear modeling. Collaborated on research utilizing advanced statistical methods to explore the relationship between teacher feelings of preparedness to teach mathematics and mathematics achievement.

Graduate Research Assistant **2009-2010**
National Research Center for Career and Technical Education
University of Louisville

Served as a research assistant on a two year project examining the prevalence of online career and technical education programs in two-year colleges. Assisted in database construction and served as lead statistician.

Research Intern **2008-2009**
Peabody College
Vanderbilt University

Collected classroom observation data for a National Science Foundation funded five-year study examining how induction and continuing professional development affect middle school mathematics teacher's instruction and student achievement.

Research Intern **2008-2009**

Jefferson County Public Schools:
Accountability, Research, and Planning Department
Louisville, Kentucky

Worked under the supervision of evaluation specialists in assisting with various research and evaluation projects. Duties included: development of literature reviews and program reports, statistical analysis of data, creation and maintenance of databases.

Administrative Assistant/ Webmaster **2005-2007**

Kentucky Psychoanalytic Institute
Louisville, Kentucky

Created the institute's promotional materials. Responsible for the institute's financial records, tuition billing, and annual financial reports. Designed and built institute's website. Supervised technology upgrade and introduction of distance learning. Served on several committees of the Institute's Board of Directors

Family Preservation Therapist **2006-2007**

Kentucky River Foothills Community Action Partnership
Shelbyville, Kentucky

Provided intensive in-home therapy for at risk families. Assessed family and developed goals based on deficits. Implemented safety plan and goal related work. Evaluated strength improvement and made placement recommendations. Served as an advocate for families at court hearings.

Research Intern
Kentucky Data Center
Louisville, Kentucky

Analyzed Kentucky murder data from 1969-2004, examining relationship between victim and perpetrator race and the chronological trends in clearance rates. Created four-page profiles of census data for Kentucky Metropolitan Statistical Areas.

Student Research Assistant
Criminal Justice Department
Bellarmino University
Louisville, Kentucky

Assisted in the development of the methodology of a proposed federally funded pilot study of a decision making tool for probation and parole officers.

UNIVERSITY COURSES TAUGHT

University of Louisville Courses. Hierarchical Linear Modeling (TA), Structural Equation Modeling(TA).

HONORS & AWARDS

Fellowships, Scholarships, & Grants

Research and Faculty Development Grant, University of Louisville, College of Education and Human Development, 2010-2011. *Examining Fourth Grade Teachers' Feelings of Preparedness in Mathematics Content Areas*, \$2,916.
Horrigan Scholarship, Bellarmine University, 2002-2006.

Norton Healthcare Scholarship, 2002.

Presidential Scholarship, Bellarmine University, 2003-2006.

Honor Societies

Kappa Gamma Pi (National Catholic College Graduate Honor Society)

Psi Chi (Psychology National Honor Society)

Omicron Delta Kappa (National Leadership Honor Society)

PUBLICATIONS

Journal Articles in Preparation, Revision, or Review

Githens, R. P., **Sauer, T. M.**, Wilson, K. B., & Crawford, F. (in press). Online occupational education in community colleges: Prevalence, programming, and connection with workforce development needs.

Adelson, J.L., & **Sauer, T.M.** (2011). *Examining Fourth Grade Teachers' Feelings of Preparedness in Mathematics Content Areas*. Manuscript in preparation.

Adelson, J.L., & **Sauer, T.M.** (2011). *Professional Development and Teachers' Feelings of Preparedness to Teach Mathematics Content Strands: A Multi-Group Analysis of TIMSS Data*. Manuscript in preparation.

Adelson, J. L., McCoach, D. B., Rogers, H. J., Adelson, J. A., & **Sauer, T.M.** (2010). Using the propensity score to make causal inferences: Variable selection and stratification issues. Manuscript in revision.

Refereed Monographs

Githens, R. P., Crawford, F., & **Sauer, T. M.** (2010). *Online occupational education in community colleges: Prevalence and contextual factors*. Louisville, KY: National Research Center for Career and Technical Education.

Non-Refereed Journal Articles

Sauer, T. M. & Githens, R. P. (2010). Going green online: Distance learning prepares students for success in green-collar job markets. *Community College Journal*, 80(6), 30-33.

Other Publications

Adelson, J. L., **Sauer, T. M.**, & Canada, D. D. (June, 2011). Girls on the Run Louisville: Spring pilot evaluation results. Report for Girls on the Run Louisville

Adelson, J. L., **Sauer, T. M.**, & Canada, D. D. (June, 2011). Body image, self-esteem, and attitudes toward physical activity for elementary school girls in the spring. Report for Girls on the Run Louisville.

PRESENTATIONS

Adelson, J.L., **Sauer, T.M.**, & Truitt, T. (2012). *Professional Development and Teachers' Feelings of Preparedness to Teach Mathematics: A Multi-Group Analysis Using TIMSS*. American Educational Research Association Annual Meeting, Vancouver, British Columbia, Canada.

Adelson, J.L., **Sauer, T.M.**, & Truitt, T. (2012). *Teachers Feelings of Preparedness to Teach Mathematics, Instructional Time, and Student Achievement*. American Educational Research Association Annual Meeting, Vancouver, British Columbia, Canada.

Wilson, K.B., Githens, R.P., Cumberland, D., & **Sauer, T.M.** (2011). *How Leaders Make Sense of Online Technical Program Development*. 53rd Annual Meeting of The Council for the Study of Community Colleges: New Orleans, Louisiana.

Sauer, T.M., Choi, N. & Carpenter, B. (2010) *A Reliability Generalization Study of the Personal Attributes Questionnaire*. American Educational Research Association Annual Meeting, Denver, Colorado.

Githens, R.P., Crawford, F. & **Sauer, T.M.** (2010) *Online occupational education in community colleges: Prevalence and contextual factors*. American Educational Research Association Annual Meeting: Denver, Colorado.

Githens, R.P., Wilson, K.B., Crawford, F.L., **Sauer, T.M.**, & Cumberland, D. (2010). *Online Contract Training in Community and Technical Colleges: A Cross-Case Analysis of Five Institutions*. 2010 ACTE Annual Convention & Career Tech Expo: Las Vegas, Nevada.

Kotamaraju, P., Blackman, O. & **Sauer, T.M.** (2010) *A Cross-State Comparison of Postsecondary Student Graduation Rates: A Macro Perspective of Educational Attainment*. Council for the Study of Community Colleges Annual Meeting: Seattle, Washington.

- Githens, R.P., Crawford, F. & **Sauer, T.M.** (2009) *Online Workforce Development: Contours and Directions*. 34th Annual Conference of the National Council for Workforce Education: Seattle, Washington.
- Sauer, T.M.** & Dossett, D. (2009) *Turning Around Schools: A Review of What Works from the Research and Reality*. Consortium for Research on Educational Accountability and Teacher Evaluation: National Evaluation Institute: Louisville, Kentucky.
- Cunningham, N., Robbins, T., DeLaney, N., Whitten, M., Eder, K., Taylor, M., Goodwin, S., Mizelle, D., **Sauer, T.M.** (2008). *Classroom Intervention to Address Relational and Sexual Bullying in Early Adolescence: A Pilot Study*. American Psychological Association Annual Convention: Boston, Massachusetts.
- DeLaney, N. & **Sauer, T.M.** (2008) *Gender differences in cyberbullying behavior among middle school students*. Kentucky Psychological Association Spring Academic Conference. Louisville, Kentucky.
- Sauer, T.M.** (2006) *An analysis of conformity and authority*. Butler University Undergraduate Research Conference: Indianapolis, Indiana.
- Sauer, T.M.** (2005) *Assessing the violation response matrix*. Bellarmine University Undergraduate Research Celebration: Speech Session: Louisville, Kentucky.
- Sauer, T.M.**, Kochersberger, J., & Wilburn, K. (2005) *A correlational study on self-esteem & locus of control*. Mid American Undergraduate Psychology Research Conference. Franklin, Indiana.