Student Ratings of Instruction:
Examining the Role of Academic Field, Course Level, and Class Size

Anne M. Laughlin

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Educational Leadership and Policy Studies

Steven M. Janosik, Chair

Yasuo Miyazaki, Co-Chair

Martha J. Glass

Claire K. Robbins

March 21, 2014

Blacksburg, Virginia

Keywords: Student Evaluation of Teaching, Student Ratings of Instruction, Holland's Theory of Academic Environments, Bootstrap, Robust Methods

Student Ratings of Instruction:
Examining the Role of Academic Field, Course Level, and Class Size

Anne M. Laughlin

Abstract

This dissertation investigated the relationship between course characteristics and student ratings of instruction at a large research intensive university. Specifically, it examined the extent to which academic field, course level, and class size were associated with variation in mean class ratings. Past research consistently identifies differences between student ratings in different academic fields, but offers no unifying conceptual framework for the definition or categorization of academic fields. Therefore, two different approaches to categorizing classes into academic fields were compared – one based on the institution's own academic college system and one based on Holland's (1997) theory of academic environments.

Because the data violated assumptions of normality and homogeneity of variance, traditional ANOVA procedures were followed by post-hoc analyses using bootstrapping to more accurately estimate standard errors and confidence intervals. Bootstrapping was also used to determine the statistical significance of a difference between the effect sizes of academic college and Holland environment, a situation for which traditional statistical tests have not been developed.

Findings replicate the general pattern of academic field differences found in prior research on student ratings and offer several unique contributions. They confirm the value of institution-specific approaches to defining academic fields and also indicate that Holland's theory of academic environments may be a useful conceptual framework for making sense of academic field differences in student ratings. Building on past studies that reported differences in mean ratings across academic fields, this study describes differences in the variance of ratings across academic fields. Finally, this study shows that class size and course level may impact student ratings differently – in terms of interaction effects and magnitude of effects – depending on the academic field of the course.

Acknowledgements

While undertaking this study, I benefited from the guidance and support provided by numerous individuals.

I am grateful for the support provided by each member of my excellent dissertation committee. Steve Janosik for meeting me where I was, staying beside me along the way, and providing the structure and encouragement to help me finish. Yasuo Miyazaki for expert advice and engaging discussions that helped me to work through the many methodological options and analytical procedures. Claire Robbins for insightful comments during the proposal stage that helped to expand my thinking about prior research. Martha Glass for thoughtful questions and warm encouragement after agreeing to join my committee just one month prior to the prospectus.

Past and present colleagues in the Office of Assessment and Evaluation – especially David Kniola and Kate McConnell, who were classmates as well – bolstered me with reassurance and humor.

My parents, James and Barbara Laughlin, nudged me forward with love. They encouraged me to begin the journey toward this degree, and strengthened my resolve to complete it by never doubting that I would.

Last but not least, the completion of this dissertation would not have been possible without the support of my husband, Jeff Busche, who provided sustaining love and encouragement every step of the way.

Table of Contents

List of Tables

List of Figures

Chapter One

Introduction

Student ratings of instruction are the most commonly used gauge of teaching effectiveness in American higher education. In a survey of more than 40,000 college and university department chairs, more than 97% indicated that they used "student evaluations" to assess teaching performance (U.S. Department of Education, 1991). Smaller numbers reported that they used other data sources, including: department chair evaluations (84%), peer evaluations (53%), self evaluations (47%), and dean evaluations (40%). In a separate survey of college deans, one wrote, "Student feedback is the most important factor in our determination of teaching effectiveness"; and another commented, "If I trust one source of data on teaching performance, I trust the students" (Seldin, 1999, p. 15).

Experts on faculty evaluation have consistently insisted on the importance of using multiple data sources for decisions about teaching effectiveness (Abrami, 2001a; Braskamp & Ory, 1994; Centra, 1993; Seldin, 1999). These authors argue that any summative decisions related to teaching quality, such as for faculty promotion and tenure, should use more than student perceptions. Still, Seldin (1999) has noted that many academic administrators rely heavily on student rating data as the only information systematically collected to evaluate teaching. Because of the widespread use of student ratings, and because decisions based on these ratings impact academic careers and the quality of instruction, those involved in higher education need to be concerned with the integrity of both the student ratings data and with how these data are interpreted and used.

This chapter provides a very brief introduction to some of what is known, and not known, about student ratings of instruction (here called "student ratings", or just "ratings"). I begin by

highlighting a few studies pulled from the large body of research that indicates student ratings are generally reliable, valid, and useful sources of data about teaching effectiveness in higher education. Next, I explain why the validity of student ratings as a measure of teaching quality is only part of the story, and why we need to expand what is known about the appropriate interpretation and use of student rating data, particularly in comparative judgments. I then explain how my study relates to the open question of how to establish appropriate reference groups for comparative judgments.

I end the chapter with information about three variables that are the focus of this quantitative study - academic field, course level, and class size. In addition to defining academic fields based on a traditional academic college structure, this study will use Holland's theory of academic environments. Holland's theory has proven useful for understanding differences in teaching preference among college faculty and differences in college student development and learning, but it has not been used in research on student ratings of instruction. This study is the first to test the theory's usefulness as a conceptual framework for explaining academic field differences in student ratings. The study compares results within and across academic fields to find potential interaction effects with course level and class size. Both course level and class size have been identified in past research as course characteristics that may need to be accounted for when interpreting student ratings data.

Student Ratings as a Measure of Instructional Effectiveness

Many college and university faculty members have questions about the trustworthiness of student ratings. They may acknowledge that student ratings are useful for the individual instructor to make pedagogical and curricular improvements, but maintain that student ratings are flawed in ways that make them inappropriate for use in decisions about teaching

effectiveness. In response to these concerns, researchers have made student ratings of instruction the focus of considerable inquiry. Multiple reviews of this large body of literature indicate that, with a few exceptions, the most common concerns about the validity of student ratings are not supported by research. The following paragraphs briefly indicate how a few of the most frequently asked questions have been answered in the research literature. Findings related to the reliability and validity of student ratings are summarized in this chapter to provide context for the problem statement and purpose. A fuller account of the related literature, including studies related to potential bias and interpretation of results, is presented in the second chapter.

Student Qualifications

The conclusion presented in multiple research reviews is that students can provide useful feedback across multiple dimensions of teaching, and that student ratings correlate well with other indicators of teaching quality - such as how much students learn, observations from experts, and alumni ratings (Cashin, 2003; Marsh, 2007). While students are not in a position to judge an instructor's content knowledge or the quality of curriculum, they are in a good position to rate and comment on other aspects of teaching - such as the amount and difficulty of work required, the value of readings, and the instructor's availably and clarity of explanation. Simply put, students are the most direct and consistent observers of an instructor's teaching (Arreola, 2007; Cashin, 1995, 1999; Feldman, 2007).

Reliability of Student Ratings

Reliability here refers to the consistency and stability of results. Consistency is most often represented in terms of the correlation between ratings within a given class (i.e., whether students in the same class tend to provide similar ratings on a given item). The consistency of any one item varies with the number of raters. Generally, the more raters there are, the more

consistent the data; however, class ratings based on as few as 10 students may still demonstrate acceptable consistency. For example, the average split-half reliabilities[1] for one well designed instrument (developed by individuals knowledgeable about educational measurement) are .78 with 10 -14 raters; .87 with 15 - 34 raters; and .92 with 35 – 49 raters (Hoyt & Lee, 2002a). A review of other studies of student ratings found that split-half reliability coefficients vary from .70 with 10 raters to .90 with 40 or more raters (Sixbury & Cashin, 1995). Scholars interpreting these results generally suggest that class averages based on 10 or more ratings are adequately reliable.

Stability refers to agreement between ratings of the same instructor across time. In general, ratings of the same instructor across semesters tend to be similar (Braskamp & Ory, 1994; Centra, 1993). For example, a longitudinal study that compared ratings given at the end of a course with ratings given by the same students a year or more later (at least a year after graduation) found an average correlation of .83 (Overall & Marsh, 1980). Another study examined the extent to which ratings measure an instructor's general teaching effectiveness, rather than just how effective the instructor was in teaching a particular class in a given term (Marsh, 2007). To distinguish the effects of instructor and course, student ratings from 1,364 classes were separated into four categories: (a) the same instructor teaching the same course but in different terms, (b) the same instructor teaching a different course, (c) different instructors teaching the same course, and (d) different instructors teaching different courses. Student ratings in the four different categories were correlated. For instructor-related items (e.g., ratings of the instructor's enthusiasm, organization, clarity), the correlations are highest for the same

[1] A measure of internal consistency in which items are randomly divided into two groups and results from each half are correlated with each other. Convention for survey research sets acceptable correlations at .7 or greater (Nunnally, 1978).

instructor, even across different courses. For student motivation items (measures tied to a particular student in a given course), the correlations are highest for the same course, even across different instructors. These results indicate that student ratings reflect an instructor's general teaching effectiveness. Additionally, they suggest that the instructor – not the course – is the primary determinant of students' ratings. The findings presented here are consistent with other reliability studies. As a whole, the research supports the reliability of a wide variety of student rating instruments.

Validity of Student Ratings

A number of researchers have explored the validity of student ratings—or the degree to which they measure what they are intended to measure—by comparing them to other indicators of effective teaching. For example, Cohen (1981) and Feldman (1989b, 2007) examined the construct validity of student ratings by reviewing studies that correlated ratings with student achievement. They focused on studies that compared multiple instructors teaching different sections of the same course. Instructors in these studies used the same syllabus, textbook, and externally developed final exam. Student ratings from each section were then correlated with the final exam scores. Both Cohen and Feldman found that classes in which the students gave the instructor higher ratings tended to be the ones where the students scored higher on the externally developed exam. A more recent example comes from the work of Benton and colleagues (Benton, Duchon, & Pallett, 2011), who correlated student ratings of progress with the students' performance on exams. On objectives the instructors deemed relevant to the course, student ratings of progress correlated positively with exam scores ($r = .32$, $p < .001$). In contrast, student ratings of progress on objects the instructors considered of minor or no importance were not related to the students' exam performance.

Researchers have also compared student ratings to the ratings provided by others in a position to judge effective teaching, such as faculty colleagues, alumni, and trained observers. In these studies, the findings indicate that student ratings correlate positively with the ratings provided by other observers. For example, a few studies have used external observers who were trained to make classroom observations (Feldman, 1989a; Marsh & Dunkin, 1992). A review of five of these studies reported an average correlation of .50 between student ratings and the ratings of trained observers (Feldman, 1989b). Other studies have compared student ratings from the end of a term with retrospective ratings provided by the same students several years later. Correlations of .54 to .80 were reported in one study (Braskamp & Ory, 1994). Another study reported average correlations of .83 (Marsh, 1982), while a third reported average correlations of .69 (Feldman, 1989b). Each indicates that students' end of term ratings are consistent with the views they hold about teaching after they graduate. In sum, these findings – all statistically significant at the .05 level – support the validity of student ratings when compared to other criteria of effective teaching (Abrami, 2001b; Cashin, 1995; Centra, 2009).

Use of Student Ratings Data

The proper use of student ratings data is as important as how it is collected. Although there is wide variability in how promotion and tenure committees use student ratings, the literature on faculty evaluation indicates that committee members typically receive little if any guidance on interpreting results and are generally ill equipped to handle the mass of data provided to them (Abrami, 2001a). Recognizing this, a few authors have offered recommendations for improving the interpretation and use of ratings for judgments about teaching effectiveness. According to Cashin (1999, p. 40), "recommendations concerning how the [student rating] data are interpreted may be the most important of all the recommendations,

but this area is the least studied." Similarly, Ory and Ryan (2001) contended that greater attention needs to be directed toward how student ratings are interpreted and used – what they call the *consequential validity* of ratings (based on Messick, 1995). They noted that the way student ratings are used "directly influences our ability to view them as appropriate, trustworthy, and credible measures of instructional quality" (p. 41).

Some have argued that one of the most effective ways to improve judgments based on student ratings is to improve the way results are summarized and reported. According to Abrami (Abrami, 2001a), "if we provide promotion and tenure committees with evidence that takes into account the general influence of extraneous factors [on student ratings], more correct decisions will be reached" (p. 101). Acting on this notion involves deciding which extraneous factors, if any, might influence the results, and making determinations about appropriate comparative data.

Comparative data provide context for interpretations and enhance the meaning and relevance of results. When individuals judge instructional effectiveness or quality, they employ either relative or absolute standards in their interpretation (Braskamp & Ory, 1994). That is, they compare one instructor's performance against the performance of others (relative comparison) or against some established standard or criteria derived through logic and/or previous experience (absolute comparison). When making either kind of comparison, a reference group (also called a norm group) influences interpretation of the ratings received by a given instructor or class. When comparisons are relative, the reference group is used to determine the relative standing of any one instructor or class. For example, an average rating of 4.8 on a 6-point scale is a favorable rating in the absolute sense, but the degree of favorability is clarified when the rating is compared to a norm group with 4.9 as its average. When absolute comparisons are made, appropriate reference groups inform the establishment of reasonable standards or criteria. For

example, analysis of the comparability of results could indicate that undergraduate and graduate courses, chemistry and biology courses, or lecture and laboratory courses, should be appraised against different absolute standards.

Cashin (1999), and Benton and Cashin (2012), have provided recommendations related to the establishment of appropriate reference groups. They noted three course characteristics that potentially bias student ratings and may require control in the form of appropriate comparative data: the academic field of the class (math and science courses receive lower ratings), course level (lower level classes receive lower ratings), and class size (larger classes receive lower ratings). Although these variables have been found to influence ratings systematically, more research is needed to address several gaps in the literature. The proposed study will begin the process by examining the relationship between course characteristics and student ratings of instruction.

Academic Fields

Ratings differ between academic fields, although it is not clear why. Humanities and arts courses receive higher ratings than social science courses, which in turn receive higher ratings than math and science courses (Barnes & Barnes, 1993; Braskamp & Ory, 1994; Cashin, 1990b; Centra, 1993, 2009; Feldman, 1978; Hoyt & Lee, 2002b; Marsh & Dunkin, 1992; Sixbury & Cashin, 1995; Smith & Cranton, 1992). Academic field differences remain even after controlling for student motivation and class size (Cashin, 1990b). Researchers have speculated that this may be because courses in math and science oriented fields are more poorly taught, or because courses in these fields require more quantitative reasoning and students are less competent in such skills (Cashin, 1990b; Feldman, 1978). It may also be because quantitatively oriented courses are more difficult to teach well.

One explanation for differences in student ratings across academic fields is based on evidence that college faculty members tend to teach in ways that reflect the culture of their academic discipline. A large body of research suggests that both what and how faculty teach is shaped by disciplinary cultures. As Becher (1981, p. 109) writes, “academic disciplines are a cultural phenomenon: they are embodied in collections of like-minded people, each with their own codes of conduct, sets of values, and distinctive intellectual tastes.” Faculty identification with the culture of their academic discipline is often stronger than with the institution where they work (Clark, 1983; Ruscio, 1987). This identification includes a discipline’s shared goals, values, and beliefs related to teaching (Becher & Trowler, 2001; Umbach, 2007). It may be that particular disciplinary cultures support more effective pedagogical approaches, resulting in higher student ratings.

Currently, two different models for classifying academic fields predominate in the research on college faculty. One is the disciplinary classification system developed by Biglan (1973) and later extended by Becher (1987; Becher & Trowler, 2001). The other is the theory of academic environments developed by Holland (1966, 1997). Both of these models have been used to help researchers make sense of the values, beliefs, and attitudes of disciplinary cultures in higher education. They continue to be cited frequently in current research and used by practitioners in student affairs and higher education administration.

Biglan’s Disciplinary Clusters

Biglan (1973) developed his classification system after asking more than 200 college and university faculty members to sort academic disciplines based on what they perceived as similarities between areas. He found that academic disciplines varied along three dimensions: hard versus soft, pure versus applied, and life versus non-life. Biglan created categories based on

the clustering of subject matter along the three dimensions in his model. These categories indicate that his participants grouped knowledge into distinct domains. *Hard* disciplines (e.g., engineering, chemistry) have a single paradigm that allow scholars within the discipline to agree on research methodology, basic concepts, and research questions. *Soft* disciplines (e.g., psychology, sociology) lack a common paradigm and often scholars within these disciplines argue about methodology and key concepts. *Pure* disciplines (e.g., mathematics, physics) focus on theory building and concepts are typically a work in progress. *Applied* disciplines (e.g., finance, education) focus on theory application and tentativeness is overshadowed by the practical need to make decisions in the field. *Life* disciplines (e.g., agriculture, biology) are concerned with living or organic objects. *Nonlife* disciplines (e.g., geology, computer science) have the absence of biological objects of study.

A few studies have applied Biglan's (1973) model to the study of student ratings of instruction and found differences in ratings along the hard/soft dimension, but not along the pure/applied or life/non-life dimensions (Barnes & Barnes, 1993; L. Neumann & Neumann, 1985; Y. Neumann & Neumann, 1983). The most recent study to examine student ratings using the Biglan model was conducted by Cashin and Downey (1995). These researchers tested whether the model could explain differences in student ratings across eight academic fields chosen to represent each of Biglan's eight disciplinary clusters. They found that students did rate the academic fields differently, although the differences were not explained by Biglan's disciplinary clusters. Based on the research to date, it seems that Biglan's theory may not be the most useful framework for making sense of contemporary differences in student ratings of academic fields. This may be because of ongoing shifts in disciplinary boundaries, increases in the number of interdisciplinary programs and departments, and/or because Biglan's dimensions

focus on research paradigms and methodology rather than on teaching.

Holland's Theory of Academic Environments

Holland's (1966, 1997) person-environment fit theory as applied in higher education settings has three basic premises, each of which is supported by a large body of empirical evidence (Holland, 1997; Smart, Feldman, & Ethington, 2000; Spokane, Meir, & Catalano, 2000; Tsabari, Tziner, & Meir, 2005). First, the choice of a career or field of training is an expression of one's personality and most people can be classified by their resemblance to six personality types (Realistic, Investigative, Artistic, Social, Enterprising, Conventional) based on their distinct patterns of attitudes, interests, and abilities. Second, there are six corresponding academic environments, each dominated by their analogous personality type. Third, the extent to which members' attitudes, interests, and abilities are congruent with their academic environments is related to higher levels of stability, satisfaction, and achievement.

In recent years, higher education scholars have increasingly used Holland's theory to make sense of patterns in how faculty members structure their courses and in what students learn as a result of their college experience. Researchers have found that faculty members use educational practices that reinforce and reward particular patterns of student competencies that are consistent with Holland's theory (Smart & Umbach, 2007). Additionally, the different educational practices favored by faculty in Holland's six academic environments are associated with varying levels of student learning and development (Astin, 1993; Kuh, 2001; Pascarella & Terenzini, 2005; Smart & Umbach, 2007). For example, faculty members in Social, and to a lesser extent Artistic, academic environments are more likely than their peers in other environments to engage students in effective educational practices such as diversity-related activities (Milem & Umbach, 2003; Umbach & Milem, 2004) and active and collaborative

learning (Umbach, 2006). More about the research that ties Holland's theory to teaching and learning in higher education is presented in Chapter Two.

Class Size

In addition to academic field, researchers have examined several other course characteristics to determine their association with student ratings of instruction. Based on findings from these studies, the size of the class is a factor that has been found sometimes to influence ratings in a way that may require control. Class size has been found to be significantly related to ratings in some studies, but not in others. When the relationship is significant, it is negative—small classes receive higher ratings than large classes (Feldman, 1984; Hoyt & Lee, 2002a). While smaller classes tend to receive higher ratings, students in those classes also score higher on self-reported learning of course outcomes (Centra, 2009). It may be that smaller classes provide a better learning environment for students and/or an environment that facilitates better teaching; either way, evidence indicates that class size may be related to both student learning and effective teaching. Consequently, researchers have concluded that class size should not be considered a biasing factor since research suggests that students in smaller classes actually do learn more. Still, class size is a variable that academic administrations may want to take into consideration by using appropriate comparative data to interpret student ratings (Benton & Cashin, 2012; Centra, 2009).

Course Level

The level of the course (e.g., upper-level versus lower-level, undergraduate versus graduate) is factor that has been found to influence ratings systematically in some studies. Although the level of the student has little effect on ratings (McKeachie, 1997), ratings in upper-level courses tend to be higher than ratings in lower-level courses (Aleamoni, 1981; Braskamp &

Ory, 1994; Feldman, 1978). More research is needed to determine whether course level impacts ratings in the same way across academic fields. For example, course level may have the greatest impact when students take a set sequence of courses that build on one another, such as is required in most engineering programs. In contrast, course level may have little impact in art and humanities programs where students have more flexibility in class selection and in the order in which courses are taken. Regarding possible control, Seldin (1999) and Davis (2009) recommend that institutional administrations determine whether their student ratings differ by course level before establishing local comparative data.

Statement of the Problem

Studies that examine academic fields have largely been based on ratings collected at liberal arts colleges, community colleges, and universities where the highest degree offered is the master's degree. Few studies have considered whether the same trends (i.e., relatively lower ratings in science and mathematics courses) are apparent in ratings collected at research universities, where more students pursue majors in science, technology, and engineering. Also, only one study of student ratings has reported on the interaction effects between academic field, course level, and class size. Finally, while past research consistently identifies differences between ratings in different academic fields, it does so outside of a unifying conceptual framework for the definition and categorization of academic fields. A conceptual framework would be beneficial because it would provide a lens for comparing research results across studies and institutions and would allow researchers to focus less on describing the influence of each academic area and more on making sense of the larger patterns. A conceptual framework could also provide a guide for college and university faculty who interpret student ratings within the context of their particular academic departments and colleges.

These gaps in the knowledge base undermine the use of student ratings in significant ways because faculty and administrators must speculate about the extent to which they can reasonably compare ratings across different academic fields, course levels, and class sizes. More research is needed to address the question of how to best account for potentially extraneous course-related variables when establishing reference groups for the interpretation of ratings. This study begins to address these issues by examining student ratings data collected at a large research institution. The mean rating for a class section was the unit of analysis because this level of aggregation is often used when student ratings factor into college and university personnel decisions.

Purpose and Research Questions

The purpose of this study was to examine the relationship between course characteristics and student ratings of instruction. Specifically, I investigated to what extent the academic field, level, and size of a class were associated with variation in mean class ratings. Two different approaches to categorizing courses into academic fields were compared – one based on the institution's own academic college system and one based on Holland's (1997) theory of academic environments. The following questions guided the study.

1. Are differences in ratings associated with the University's academic colleges?
2. Are differences in ratings associated with Holland's academic environments?
3. Are differences in ratings associated with course level?
4. Are differences in ratings associated with class size?
5. Are there interaction effects between the University's academic colleges and course level or class size?
6. Are there interaction effects between Holland's academic environments and course level

or class size?

7. Which academic fields result in more internally consistent groups: those based on the University's academic college system or those based on Holland's academic environments?

I began to help to answer these questions by using analysis of variance to examine student ratings of instruction data collected from more than 4,000 class sections at a large research university.

Significance of the Study

The results of this study will be significant for future practice, research, and policy in higher education. In terms of practice, faculty members may benefit from improved information about appropriate and inappropriate comparisons between student ratings of different classes. One way that faculty members make sense of the student ratings they receive is by comparing them to ratings received by colleagues in their department, college, and institution. However, some of these comparisons may not be justified if factors outside the instructor's control have a significant impact on the ratings. Knowing which course characteristics tend to influence class averages may help faculty members to interpret student ratings in more appropriate contexts and guide teaching improvements.

Members of promotion, tenure, and merit committees may find the results valuable when using student ratings as the basis for summative judgments about teaching effectiveness. Information related to appropriate comparative data for the interpretation of ratings can increase the credibility and trustworthiness of student ratings as a measure of teaching quality. This is particularly important when ratings are used in decisions that affect academic careers and shape academic departments.

Better information about appropriate comparison groups for the interpretation of ratings may also lead to better policies about student ratings use. For example the results may be used by academic administrators, who could apply the findings when developing policies related to the use of student ratings in promotion and tenure decisions. Individuals who manage student ratings systems for colleges and universities may also use this study's findings when developing policies for reporting student ratings to internal and external stakeholders.

The study results may also inform research on student ratings by testing Holland's theoretical framework for the classification of academic fields. If researchers can organize academic fields according to a common theoretical framework it will facilitate cross-institution comparisons by defining like-fields outside of the academic structure of any one institution. Future research may also build on the findings from this study by testing the usefulness of Holland's framework for explaining academic field differences in student ratings at different institutions and institutional types.

Organization of the Study

This study is organized in five chapters. Chapter One introduced the study's topic and context, along with the problem, purpose, and research questions. Chapter Two provides a review of the literature on student ratings of instruction with a focus on the theoretical model and variables to be examined in the current study. Chapter Three presents the methods for the study. Results of the study are presented in Chapter Four, and a discussion of results and implications are presented in Chapter Five.

Chapter Two

Literature Review

References to bias are common in the literature on student rating (Abrami, 2001b; Centra, 2003; Feldman, 2007; Marsh, 2007). Some authors use the term bias to refer to anything that systematically influences student ratings and is not under the control of the instructor. Marsh (2007) narrows the definition considerably - arguing that bias exists "when a student, teacher, or course characteristic affects the evaluations made, either positively or negatively, but is unrelated to any criteria of good teaching" (p. 350). According to this definition, correlations between ratings and class size or between ratings and student motivation are not biases because research suggests that students in small classes and students who are interested in the subject matter do tend to learn more. However, this leaves the question of how to account for such variables most appropriately when interpreting student ratings. Rather than use the term "bias," Benton and Cashin (2012) distinguish between variables (when correlated with student ratings) that may require control versus those that do not require control. This perspective is consistent with the approach I take in the present study.

Although course variables are the focus of this study, research related to instructor and student variables is also summarized in this chapter because it provides context for understanding the role of course variables. To the extent that instructor, student, and course variables impact student ratings, they do so at different levels of influence. That is, student characteristics may be expected to predominantly influence the variability of ratings between students, while instructor and course characteristics may be expected to predominantly influence the variability of ratings between classes. In the literature on student ratings of instruction, research designs tend to reflect these differing levels of influence in their unit of analysis. The

present study uses the mean rating for a class section as the unit of analysis because it is often used when student ratings factor into personnel decisions.

Variables That Do Not Require Control

Researchers have found relatively few variables that correlate with student ratings and are not related to instructional effectiveness or student learning. Studies have found that the following variables are not related to student ratings in ways that require control: instructor age, teaching experience, gender, and personal characteristics; student age, gender, GPA, and expected grades; course time of day, and time during the term (Benton & Cashin, 2012).

Instructor Variables: Age, Teaching Experience, Gender, Personal Characteristics

Instructor age and years of teaching experience are generally not correlated with student ratings. In a couple of studies where correlations were found, they are weak and negative (i.e., older instructors receive lower ratings) (Feldman, 1983; Renaud & Murray, 1996). After noting that most studies of these variables have been cross-sectional comparisons of faculty cohorts that represent different age groups, Marsh and Hocevar (1991) conducted a longitudinal study of student ratings of the same instructors across 13 years. They found no systematic changes based on age within instructors across time.

The gender of the instructor does not affect student ratings in a meaningful way (Centra & Gaubatz, 2000). Feldman (1992) conducted a review of 14 experimental studies (students rated descriptions of fictitious instructors who varied on gender) and found few gender differences in ratings. Feldman (1993) also conducted a review of 28 studies involving student ratings of real instructors and found a very weak average correlation between instructor gender and student ratings that favored female instructors. Women received slightly higher ratings on sensitivity and on concern with student level of preparedness and progress, although ratings did

not differ meaningfully between genders on other dimensions of teaching. Centra (2009) also found that female instructors received slightly higher ratings than male instructors, but he concluded that gender was not necessarily a bias because the higher ratings were also associated with differences in teaching style (i.e., female instructors were more likely than male instructors to use discussion rather than lecture, and to be more nurturing and student-centered). Benton and Cashin (2012) reviewed the studies listed above and argued that the effects associated with gender are so small that they would likely not impact summative judgments about teaching quality.

Few instructor personality traits correlate with student ratings (Braskamp & Ory, 1994; Centra, 1993). In one study, instructor enthusiasm and positive self-esteem were the only two (out of 14) personality traits correlated with student ratings (Feldman, 1986). In another study researchers found significantly different patterns of correlations between personality traits and student ratings among psychology instructors teaching six different types of courses. The researchers concluded that instructor personalities tended to be differently suited to different types of courses (Murray, Rushton, & Paunonen, 1990).

What matters more than personality is how the instructor's personal characteristics are expressed in the classroom. Erdle, Murry, and Rushton (1985) found that most of the relationship between instructor personality and student ratings can be explained by the observable behaviors the instructor displays when teaching. Similarly, Braskamp & Ory (1994, p. 180) conclude that the effect of instructors' personalities on ratings "may be caused more by what they do in their teaching than by who they are." Benton & Cashin (2012) suggest that the personality traits associated with student ratings reflect behaviors that enhance teaching effectiveness and should, therefore, not be controlled.

Student Variables: Age, Gender, GPA, Expected Grades

Researchers have not found evidence for a meaningful relationship between ratings and the student variables of age (Centra, 1993), gender (Feldman, 1977, 1993, 2007), or GPA (Davis, 2009). However, the relationship between ratings and students' expected grades is less clear. One study of a diverse sample of more 55,000 classes reported that expected grades had no effect on ratings after controlling for class size, teaching method, and student ratings of progress on learning outcomes (Centra, 2003). However, several studies have found small to medium[2] positive correlations (.10 to .30) between class-average expected grades and ratings (Braskamp & Ory, 1994; Feldman, 1997; Howard, 1980; Marsh & Dunkin, 1992, 1997; Marsh & Roche, 2000). A few possible explanations have been offered for these positive correlations (Marsh, 2007). According to the *grading leniency explanation*, instructors who give higher-than-deserved grades receive higher-than-deserved ratings. The *validity explanation* proposes that higher expected-grades reflect greater student learning – and that a positive correlation between student learning and ratings of instruction supports the validity of the measure. The *student characteristics explanation* posits that variables such as prior subject interest may affect student learning, student grades, and teaching effectiveness – so the expected-grade effect is spurious.

Reviews of research on the relationship between expected-grades and student ratings report that most findings support the validity explanation, with some support for student characteristics (Abrami, Perry, & Leventhal, 1982; Centra, 2003; Howard, 1982; Marsh, 2007). In evaluating alternative explanations for this relationship, Marsh (2007) notes that the critical variable is grading leniency rather than expected-grades per se; however, it is difficult to isolate the effects of expected grades from grading leniency. In one of the few studies to measure

[2] Cohen's (1988, p. 83) rule of thumb for effect sizes: $|r| = .10$ is small, $|r| = .30$ is medium, $|r| = .50$ is large.

grading leniency directly (rather than expected-grades), Marsh and Overall (1979) reported that teacher and student ratings of "grading leniency" were not substantially related to global ratings of teaching effectiveness. In another study teachers who reported that they were "easy" graders received significantly lower ratings (a finding that contradicts the grading leniency explanation) (Marsh & Roche, 2000). Although the relationship continues to be contested, Theall and Franklin (2001) write "… the conclusion reached by most researchers is that there should be a relationship between ratings and grades because good teaching leads to learning, which leads to student achievement and satisfaction, and ratings simply reflect this sequence" (p. 51).

Course Variables: Time of Day, Time During Term

Student ratings are not influenced by the time of day a course is taught (Aleamoni, 1981; Feldman, 1978) or by the time during the term when ratings are collected. For example, Feldman (1979) found that ratings collected any time during the last half of the term yield similar results. Carrier, Howard, and Miller (1974) found no difference between ratings collected the last week of class versus the day of the final examination. Frey (1976) found no difference between ratings collected the last week of class versus the first week of the subsequent term.

Variables That May Require Control

The research cited thus far indicates that several variables suspected of biasing student ratings are not associated with them in meaningful ways. However, some variables do influence student ratings in ways that may need to be considered when interpreting results. No instructor characteristics have been found to bias ratings in ways that require control; however, two student characteristics (interest in the subject and perception of workload) and three course characteristics (academic field, class size, and course level) may require control. Research findings on these student and course characteristics is summarized in the next two sections.

Student Variables: Prior Interest, Perception of Workload

Prior interest in the subject matter is one of the student variables that has been found to correlate with ratings. Instructors are more likely to receive high ratings from students who enter the class with a strong interest in the subject matter (Marsh & Dunkin, 1992, 1997). For example, Hoyt and Lee (2002b) found that student agreement with the item, "I really wanted to take this course regardless of who taught it" correlates positively with student ratings of the teacher, courses, and learning objectives. Related to this correlation are findings that indicate ratings are associated with students' reason for taking a course (Centra, 2009; Marsh, 2007). Centra (2009) found that required courses may receive lower ratings than other kinds of courses, but Hoyt and Cashin (1977) found that some "required" courses are very popular with students and some "elective" courses are regarded less positively. Students' prior interest in the subject matter is therefore a more useful control variable (Benton & Cashin, 2012).

Perception of class workload is another student variable that is associated with student ratings. Contrary to many expectations, student give slightly higher ratings to challenging classes (Centra, 1993, 2003; Marsh, 2001; Marsh & Roche, 2000). However, research indicates that this relationship is not straightforward. A few studies have reported non-linear relationships between workload/difficulty and student ratings (Centra, 2003; Marsh, 2001; Marsh & Roche, 2000). For example, Centra (2003) found that classes received lower ratings when they were perceived to be either too difficult or too elementary; classes received higher ratings when the difficulty/workload was perceived to be "just right." Marsh (2001) examined the nature of perceived workload and found two uncorrelated components: *bad workload* (i.e., time spent on activities that students did not consider valuable) and *good workload* (i.e., time spent on activities that helped students learn). According to Marsh, bad workload is negatively correlated

with student ratings and good workload is positively correlated.

An implication of the research outlined previously is that interest in the subject matter and perception of course workload/difficulty are student characteristics that may need to be considered when interpreting student ratings of instruction (Benton & Cashin, 2012). One process for adjusting scores to control for the influence of these student variables has been described by Hoyt and Lee (2002a). The process uses student ratings of motivation/interest, course difficulty, and effort put forth on academic work to adjust scores after controlling for the instructor's influence (amount of reading, amount of other work, stimulating students' intellectual effort) on the difficulty of the subject matter. The process results in scores that have been adjusted to account for these extraneous influences.

Course Variables: Academic Field, Course Level, Class Size

Ratings are also associated with a few course characteristics that are not related to instructional effectiveness: academic field, class size, and course level. Research indicating a relationship between these variables and student ratings was introduced in chapter one. As a reminder, humanities and arts courses tend to receive higher ratings than social science courses, which in turn receive higher ratings than math and science courses (Braskamp & Ory, 1994; Cashin, 1990a; Centra, 1993, 2009; Feldman, 1978; Hoyt & Lee, 2002b; Marsh & Dunkin, 1992; Sixbury & Cashin, 1995). Upper-level classes receive slightly higher ratings than lower-level classes (Aleamoni, 1981; Braskamp & Ory, 1994; Feldman, 1978). Some studies have found no significant relationship between class size and ratings although others have found that ratings are higher in small classes (Centra, 2009; Feldman, 1984; Hoyt & Lee, 2002a).

Academic field differences. Although student ratings of overall teaching effectiveness are higher in some academic field than in others, the reasons for these differences are not clear.

One possible explanation arises from studies of teaching behaviors that have been independently correlated with student ratings. For example, Franklin and Theall (1992) compared teachers in humanities, business, and science and engineering in terms of instructional goals, teaching methods, and grading practices. They found that, compared to teachers in other fields, humanities teachers tended to emphasize "thought" goals more so than "fact" goals and to use discussion and independent projects rather than lecturing alone. Thought goals and discussion have both been shown to correlate positively with student ratings.

Murray and Renaud (1995) studied teaching behavior in three disciplinary groups – arts and humanities, social sciences, and natural sciences and mathematics. They observed the occurrence of specific "low inference" (i.e., directly observable) teaching behaviors that had been correlated with student ratings of instruction in prior studies (Murray, 1983, 1985). These behaviors included reviewing topics from previous classes, writing key terms on the board, using concrete examples to explain concepts, and encouraging students to ask questions. They found that arts and humanities teachers exhibited a wider range of teaching behaviors that correlate positively with student instructional ratings than did social science or natural science teachers. Furthermore, they found no difference between the disciplinary groups in the correlation of specific teaching behaviors with student ratings of effectiveness. In other words, while the specific teaching behaviors varied in frequency across disciplinary groups, the specific behaviors that made a teacher effective in the eyes of students were the same across disciplines. These findings are consistent with three previous studies that reported similar patterns of teacher behavioral difference among academic fields together with a lack of difference among academic fields in the correlation of specific teaching behaviors with student ratings of effectiveness (Erdle & Murray, 1986; Pohlmann, 1976; Solomon, 1966).

Findings on academic field differences in student ratings are difficult to synthesize because researchers define academic fields differently. In the studies just described for example, Murray and Renaud (1995) put biology, physics, anatomy, mathematics, and computer science classes in the same disciplinary group. Franklin and Theall (1992), on the other hand, had distinct categories for classes in natural sciences, health-related professions, and mathematics. Neither of the study descriptions provided a rational for the organizational framework used to categorize classes into academic fields.

The present study uses Holland's theory of academic environments (Holland, 1997; Smart et al., 2000) as a framework for categorizing classes by academic field. The university's academic college structure will provide a baseline for comparison to determine the usefulness of Holland's theory.

Holland's theory of academic environments. According to Holland's (1973) person-environment fit theory most people display patterns of attitudes, interests, and abilities that can be classified into one of six personality types. These personality types have six corresponding model environments that create a typology for academic disciplines – realistic, investigative, artistic, social, enterprising, and conventional (Smart et al., 2000). Faculty members prefer academic environments that are compatible with their personality type and in turn are rewarded for participation in activities valued by that particular environment. Greater person–environment congruence is associated with higher levels of satisfaction and success. Faculty members are thus socialized toward displaying behaviors that match their particular academic environment. Descriptions of the six environments are provided in Table 1 (adapted from Holland, 1997; Smart et al., 2000). Holland's theory has proven useful in explaining differences in teaching practices among college faculty, including differences in the use of educational practices that are

Table 1

Holland Environments and Example Disciplines

Environment	Example Disciplines
Realistic environments emphasize concrete, practical activities and the use of machines, tools, and materials. Realistic environments reward people for the display of conforming behavior and practical accomplishments.	Building Construction, Crop and Soil Sciences, Mechanical Engineering, Mining Engineering
Investigative environments emphasize analytical or intellectual activities aimed at the creation and use of knowledge. Investigative environments reward people for skepticism and persistence in problem solving, documentation of new knowledge, and understanding solutions of common problems.	Biology, Chemistry, Civil Engineering, Computer Science, Economics, Industrial Engineering, Sociology
Artistic environments emphasize ambiguous, free, and non-systemized activities that involve emotionally expressive interactions with others. Artistic environments reward people for imagination in literary, artistic, or musical accomplishments.	Architecture, Art, Communication, English, Foreign Languages, Music, Philosophy
Social environments emphasize activities that involve the mentoring, treating, healing, or teaching of others. Social environments reward people for the display of empathy, humanitarianism, sociability, and friendliness.	Education, Human Development, Area Studies
Enterprising environments emphasize activities that involve the manipulation of others to attain organizational goals or economic gain. Enterprising environments reward people for the display of initiative in the pursuit of financial or material accomplishments, dominance, and self-confidence.	Business, Finance, Hospitality, Management, Marketing

Conventional environments emphasize activities that involve the explicit, ordered, systematic manipulation of data to meet predictable organizational demands or specified standards Conventional environments reward people for the display of dependability, conformity, and organizational skills.	Bookkeeping, Data Processing, Medical Administrative Services, Retail Operations

associated with student learning and development (Astin, 1993; Kuh, 2001; Pascarella & Terenzini, 2005). For example, Smart and Umbach (2007), asked the extent to which faculty members purposefully structure their undergraduate courses to foster student learning and development in 12 different areas, such as acquiring a broad general education, acquiring work-related skills, writing, thinking critically, analyzing quantitative problems, and solving complex real-world problems. They found distinctions in how faculty from different Holland environments design and structure their classes. Further, they found that the distinctions were consistent across four-year colleges and universities representing five different Carnegie institutional types.

Researchers have examined the particular values and teaching practices associated with each Holland environment and found that faculty members in social environments interact with students more frequently, use diversity-related activities more frequently, and place greater importance on enriching actives than do faculty members in other environments. Faculty members in social environments design and structure their courses to assist students in developing the ability to recognize and seek solutions to interpersonal problems, while faculty members in enterprising environments design and structure their courses to help students acquire work-related skills (Umbach, 2006, 2007; Umbach & Wawrzynski, 2005).

Realistic, investigative and conventional faculty members are the least likely to incorporate racial and ethnic diversity issues into classroom discussions and assigned readings (Milem & Umbach, 2003; Milem, Umbach, & Liang, 2004; Umbach, 2006, 2007; Umbach & Milem, 2004; Umbach & Wawrzynski, 2005). They are also among the least likely to employ active and collaborative learning (Morstain & Smart, 1976; Peters, 1974; Umbach, 2006, 2007; Umbach & Wawrzynski, 2005). Investigative and conventional faculty members are the least

likely to emphasize higher-order cognitive activities in their classes, while faculty members in realistic environments are the most likely to emphasize these activities (Umbach, 2006, 2007; Umbach & Wawrzynski, 2005).

Interactions among academic field, course level, and class size. Only one study of student ratings has explored the question of whether there are differences between academic disciplines in terms of the amount of variation associated with class size and course level. To determine whether patterns and magnitudes of association among certain course and instructor variables varied by discipline, Franklin and Theall (1995) examined student ratings data collected in more than 8,000 course sections in a large research university. They categorized their data into eleven disciplinary groups (the rationale for their categorization was not provided). They conducted the same analysis within each disciplinary group and compared the proportion of variance in overall instructor effectiveness accounted for by each variable across disciplines. They found that relationships between class size, course level, and student ratings were "strikingly different from discipline to discipline, suggesting that these variables are not related to student ratings in the same way in each discipline" (1995, p. 43).

The findings in Franklin and Theall's (1995) study suggest that class size and course level may impact student ratings differently, depending on the academic field of the course. This indicates that the influence of class size and course level may be best understood within particular academic fields. More research is needed to clarify the nature of this interaction and to determine if class size and course level are correlated with student ratings to the same extent across different academic fields. The present study will explore this issue by analyzing variations in mean class ratings based on differences in class size, course level, and academic field. Individual class sections provide the unit of analysis, thus keeping the focus on course-

specific (rather than student-specific) variables.

Dependent Variable: Overall Teaching Effectiveness

There is wide agreement that student ratings are multidimensional, that is, they reflect several different aspects of teaching. A number of factor-analytic studies have derived these dimensions statistically (Abrami & d'Apollonia, 1990, 1991; Abrami, d'Apollonia, & Rosenfeld, 2007; Benton & Cashin, 2012; Marsh & Dunkin, 1997). The number of dimensions varies with the form being studied and the number and kind of individual items that form contains. Two published reviews have identified six factors that are commonly found in student-ratings data (Braskamp & Ory, 1994; Centra, 1993): (a) course organization and planning; (b) clarity and communication skills; (C) teacher student interaction and rapport; (d) course difficulty and workload; (e) grading and examinations; and (f) student self-rated learning. The multidimensionality found in ratings data indicates that students can distinguish among multiple factors related to teaching effectiveness and that students differentially weight various teaching behaviors when making overall evaluations of their instructors.

Observing the multidimensionality of student ratings, researchers generally agree that various dimensions should be used when the goal is to improve teaching. Teaching methods may vary depending on the course content, student characteristics, and size of class, making certain dimensions of teaching more relevant in some classes than in others. McKeachie (1997) argued that effective teaching can be demonstrated in many ways, and no instructor should be expected to demonstrate equal proficiency in all methods and styles. Similarly, Hoyt and Lee (2002b) advise instructors to distinguish among the various items on their rating form to ensure that they are attending to teaching dimensions that are relevant to their style and method of instruction.

For personnel decisions it is not necessary to distinguish among all dimensions of

teaching. Several researchers have suggested that one or a few global items should be used for personnel decisions (Abrami, 2001a; Abrami & d'Apollonia, 1991; Braskamp & Ory, 1994; Cashin & Downey, 1992; Centra, 1993). These authors point out that global items have relatively higher validity coefficients when compared to measures of specific instructional dimensions, and instructional styles and methods are likely to have larger effects on specific instructional dimensions than on global items. Also, multiple studies have found evidence of a core construct underlying the various dimensions often measured in student ratings (Abrami & d'Apollonia, 1991; Abrami et al., 2007). These findings are consistent with secondary analysis indicating that different rating items are highly inter-correlated despite their apparent conceptual independence (Abrami & d'Apollonia, 1991; Marsh, 1991).

Summary

Researchers have found relatively few variables that correlate with student ratings but are not related to teaching effectiveness or student learning. Variables such as the instructor's age and gender, the student's GPA and expected grades, and the course time of day do not influence student ratings in ways that require control. Variables that influence student ratings in ways that may require control include the student's prior interest in the subject matter and perception of workload, and the course academic field, course level, and class size.

The potential influence of academic field on student ratings is particularly difficult to understand. This is because no common framework for the definition of academic fields exists in this literature; rather, each study uses a different approach to categorizing classes into disciplines. Also, one study found significant differences between academic fields in terms of the amount of variation associated with class size and course level – this suggests that the influence of these course variables may be most appropriately observed within academic fields.

This study builds on the research literature by examining the specific course variables that research has found to be associated with student ratings. The study focuses on overall teaching effectiveness, rather than multiple dimensions of teaching, because past studies have provided evidence for the validity and usefulness of this type of global item in personnel decisions. The study will contribute to the body of knowledge by testing the usefulness of Holland's theory of academic environments as a framework for the classification of classes and by examining potential interactions between academic fields and course level or class size.

Chapter Three

Methodology

The purpose of this study was to examine the relationship between course characteristics and student ratings of instruction. Specifically, I investigated to what extent the academic field, level, and size of a class are associated with variation in mean class ratings. Two different approaches to categorizing courses into academic fields were compared - one based on the University's academic college system and one based on Holland's (1997) theory of academic environments. The following questions guided the study.

1. Are differences in ratings associated with the University's academic colleges?
2. Are differences in ratings associated with Holland's academic environments?
3. Are differences in ratings associated with course level?
4. Are differences in ratings associated with class size?
5. Are there interaction effects between the University's academic colleges and course level or class size?
6. Are there interaction effects between Holland's academic environments and course level or class size?
7. Which academic fields result in more internally consistent groups: those based on the University's academic college system or those based on Holland's academic environments?

Participants

The study used data collected from students at a large, Research I, state-supported university in the southeastern United States. The institution enrolls approximately 24,000 undergraduate, and 5,000 graduate and professional students. The Student Perceptions of

Teaching (SPOT) process provided a standardized, university-wide method for collecting student feedback regarding courses and instruction. The participants for this study were undergraduate students who responded to the SPOT questionnaire during the 2012 Fall and Spring terms. All student responses were aggregated at the level of the class section prior to analysis. The mean rating for a class section was the unit of analysis because this level of aggregation is often used when student ratings factor into college and university personnel decisions.

Role of the Researcher

The Director of the Office of Assessment and Evaluation, and the Institutional Review Board (IRB) for the Protection Human Subjects granted approval for the use of SPOT data in this study (see Appendix A). My role as Assessment Coordinator in the Office of Assessment and Evaluation involves working with SPOT data; however, this research was distinct from my job responsibilities, and my supervisor was not on my dissertation committee. I had no connection to the students who submitted the data since my study involved no intervention. My professional role does not involve teaching or advising students, and I had no opportunity to influence the individual students who already provided the existing data. The data I downloaded from the University's data warehouse were already aggregated at the level of the class section. The data set contained only the mean ratings for each class section and did not contain results submitted by individual students.

Questionnaire

The SPOT questionnaire has 12 standard items that were developed by a committee composed of representatives from the university's teaching and learning center, assessment office, and various academic departments. The items solicit student perceptions regarding aspects of instructor performance, such as organization, success in communicating subject

matter, and respect for students as individuals. In addition to the 12 standard items, each academic college may include a limited number of college-specific and/or department-specific items; however, the total number of questionnaire items associated with each class cannot exceed 20.

The following types of courses are excluded from the SPOT process by default: projects and reports, seminars, field studies, independent studies, and undergraduate research. Other courses are excluded on a case-by-case basis if, for example, the course has a non-standard academic calendar, a large team of instructors (i.e., more than three), or if the instructors rotated mid-term. Approval to exclude a course is granted by the dean or associate dean of each college. Courses that are not excluded in one of these ways are automatically included.

The online SPOT questionnaire is open to all students enrolled in participating classes for approximately two weeks at the end of each semester. Students receive an email with a link to the online questionnaire when it opens. They are then reminded via email every two days until they complete the questionnaire or the closing date (one day prior to the start of final exams) is reached.

Data for this study were downloaded from the University's data warehouse – the central repository for all SPOT results. The data file included the mean ratings for each undergraduate class that participated in the SPOT process. Each row in the data file represented one class (i.e., the unit of analysis). Columns contained data for the following variables: year, term, academic college, subject code (i.e., subject area of the course), home department, course number, section number, enrollment, response count, and class mean for each item on the SPOT questionnaire.

Analysis

The dependent variable in this study was the mean class rating on one of the standard

SPOT questionnaire items: “Overall, this instructor’s teaching was effective.” Responses to this item were given on a six-point Likert-scale (strongly disagree to strongly agree). This item was selected as the focus of study because the greatest weight is typically placed on this global rating item by promotion, tenure, and award committees and this item is of greatest concern to instructors in the performance review process.

Academic college was one of the independent variables that was examined. Six of the institution’s seven academic colleges were included in the study: Agriculture and Life Sciences, Architecture and Urban Studies, Business, Engineering, Liberal Arts and Human Sciences, and Science. The College of Natural Resources and Environment was not included in the study because exploratory analysis indicated that there were not enough class sections in this college to provide acceptable statistical power.

Academic environment as defined by Holland’s theory was another independent variable. The class subject code was used to assign each class to one of Holland’s model academic environments. The Educational Opportunities Finder (Rosen, Holmberg, & Holland, 1997) applies Holland’s theory to the classification of more than 900 college fields and was used to guide this process. To increase reliability, a second person replicated the process for assigning each class to a Holland environment. The results were compared and differences in the assignment of classes were discussed. The data set contained 75 distinct subject codes and 71 of these subjects were assigned to the same Holland environment by both individuals. The 4 subjects that were not similarly assigned represented 43 college- or department-level interdisciplinary and general studies classes. Because the 43 class sections in these 4 subjects could not be reliably assigned to a Holland environment they were removed from the data set. No subjects were assigned to Holland’s conventional environment. Appendix B shows the

categorization of subjects into colleges and environments.

Class size is a third independent variable that was examined in this study. It was based on class enrollment data. Each class was categorized in one of two levels based on the number of students enrolled: small (37 or fewer) or large (more than 37). The dividing point for small versus large classes was selected based on exploratory analysis with a subset of the data; the analysis indicated that approximately one-half of class sections fell into each of these categories.

Course level was the fourth independent variable. Each class was categorized into one of two levels based on the course number. Course numbers 1000 to 2999 were categorized as lower-division classes. Course numbers 3000 to 4999 were categorized as upper-division classes.

Table 2 shows the analyses that were used to address each research question. For research questions one through six, ANOVA was used to examine differences in the continuous variable of mean class rating based on the categorical variables of academic field, course level, and class size. ANOVA provides a way to test for differences between means and, therefore, fit the purpose of the study – which was to examine differences between ratings based on course characteristics.

For research question seven, the eta squared value from the one-way ANOVA for academic college was compared to the eta squared value from the one-way ANOVA for Holland environment. In the context of a single factor ANOVA, eta squared provides an estimate of the proportion of total variance associated with the independent variable. Thus, for purposes of this study, it is a reasonable indicator of the internal consistency of the groups.

Table 2

Research Questions and Analyses

1.	Are differences in ratings associated with the University's academic colleges?	One-way ANOVA with 6 colleges. Post-hoc pairwise comparisons.
2.	Are differences in ratings associated with Holland's academic environments?	One-way ANOVA with 5 environments**.** Post-hoc pairwise comparisons.
3.	Are differences in ratings associated with course level?	One-way ANOVA with 2 course divisions.
4.	Are differences in ratings associated with class size?	One-way ANOVA with 2 class sizes.
5.	Are there interaction effects between the University's academic colleges and course level or class size?	Three-way ANOVA (6 colleges x 2 class sizes x 2 course divisions). Two-way ANOVA (2 class sizes x 2 course divisions) in each college.
6.	Are there interaction effects between Holland's academic environments and course level or class size?	Three-way ANOVA (5 Holland environments x 2 class sizes x 2 course divisions). Two-way ANOVA (2 class sizes x 2 course divisions) in each environment.
7.	Which academic fields result in more internally consistent groups: those based on the University's academic college system or those based on Holland's academic environments?	Compare the values of eta-squared from one-way ANOVAs.

Data Cleaning and Screening

Osborne's (2013) *Best Practices in Data Cleaning* provided a guide for cleaning and examining the data prior to analysis. The process included scanning for incomplete or unreliable data, testing distributional assumptions, and calculating statistical power.

Past studies of student ratings of instruction (Hoyt & Lee, 2002a; Sixbury & Cashin, 1995) indicate that class averages based on 10 or more raters demonstrate sufficient reliability. Therefore, only class averages based on 10 or more raters were kept in the dataset for this study. The dataset was also screened for incomplete or missing scores – none were found.

The characteristics of the overall distribution were examined next. The histogram in Figure 1 shows that the distribution of ratings is highly skewed and leptokurtic. The 4,599 class ratings range from 1.76 to 6.00 with a mean of 5.03, standard deviation of .69, skew of -1.21, and kurtosis of 1.61.

The normality assumptions of ANOVA do not apply to the overall distribution of data, but to the distributions within each group. To test the tenability of these assumptions, I examined histograms and P-P plots of the ratings within each cell of the College x Course Level x Class Size design. All of the distributions deviated from normal.

Levene's (1960) test was used to check for homogeneity of variance in the College x Course Level x Class Size design. Levene's test performs a one-way ANOVA on the deviation scores to determine whether group variances are equal. The result was significant, $F(23, 4575) = 25.41$, $p < .001$, indicating that variance in ratings is not equal across the various combinations of academic college, course level, and class size.

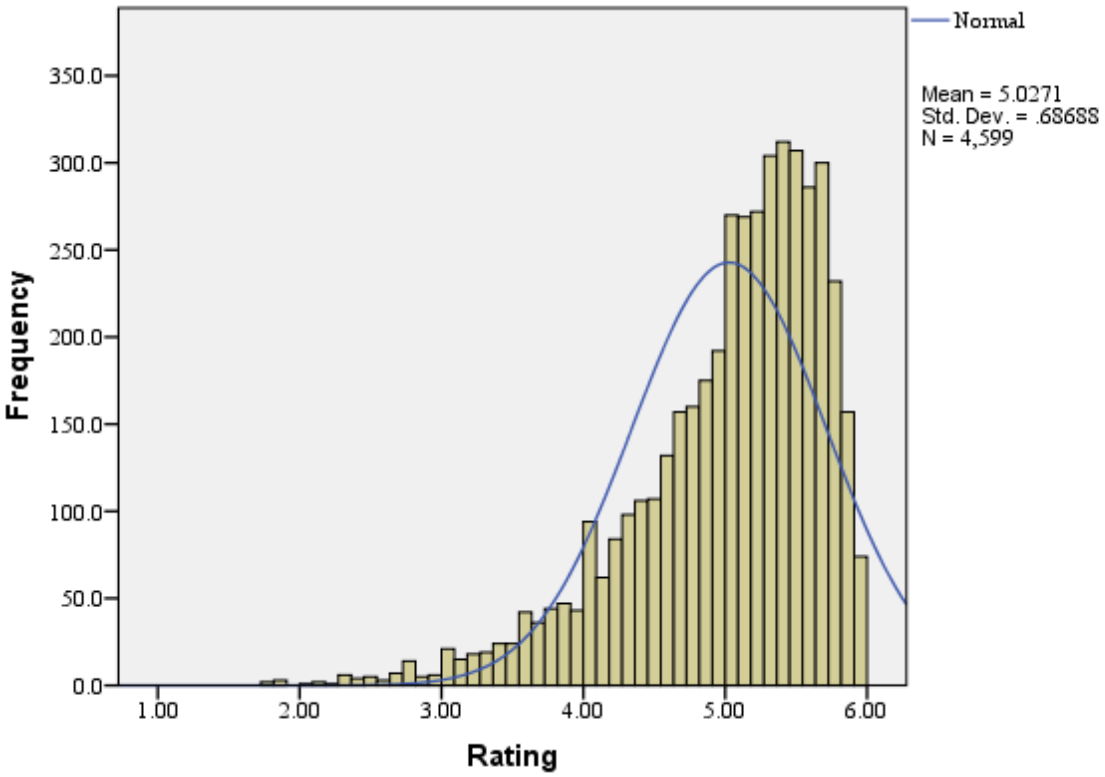

Figure 1. Overall Distribution of Ratings

A sensitivity power analysis was conducted on the College x Class Size x Course Level design[3]. This analysis determines the minimum effect size to which a statistical test will be sensitive for a given alpha level, power level, and sample size. Results indicated that a one- and three-way ANOVA would be sensitive to small effects and a two-way ANOVA would be sensitive to medium effects based on Cohen's (1988) effect size guidelines (see Appendix C for input parameters and results).

Robust Procedures

Tables 3 and 4 show that cell sizes in this study are unequal (i.e., the design is unbalanced). This is important to consider because it means that the variables of interest do not account for independent proportions of the variance, as they would in a balanced design. Put differently, the unequal *n* is problematic because cells with a larger *n* may contribute relatively more to the overall variability in ratings. To account for this I used Type III Sums of Squares to calculate the sum of squares attributable to each source of variation. Type III Sums of Squares are based on equally weighted means (also known as unweighted means, least squared means, or estimated marginal means)[4]. In the context of ANOVA, marginal means are considered to be *equally weighted* when they are calculated by taking the average of the individual cell means at a given level. This method of calculating marginal means removed the effect of unequal *n* by giving each cell equal weight in the sums of squares analysis (Roberts & Russo, 1999).

Another issue with the data used in this study is that they do not meet the assumptions of normality and homogeneity of variance. This is problematic for statistical tests that assume normal distributions within groups and equal variances across groups (especially in unbalanced

[3] Power analyses were done using GPower (version 3.1).

[4] Type III Sum of Squares is the default option in many statistical software pages, such as SPSS, SAS and STATA.

Table 3

Number of Class Ratings per Cell: College by Class Size by Course Level

		Course Level		
	Class Size	Lower	Upper	Total
Liberal Arts & Human Sciences	Small	680	440	1,120
	Large	222	80	302
	Total	902	520	1,422
Agriculture & Life Sciences	Small	85	151	236
	Large	59	57	116
	Total	144	208	352
Architecture	Small	56	96	152
	Large	21	31	52
	Total	77	127	204
Business	Small	27	163	190
	Large	73	163	236
	Total	100	326	426
Science	Small	658	172	830
	Large	462	102	564
	Total	1,120	274	1,394
Engineering	Small	220	205	425
	Large	137	239	376
	Total	357	444	801
All Colleges	Small	1,726	1,227	2,953
	Large	974	672	1,646
	Total	2,700	1,899	4,599

Table 4

Number of Class Ratings per Cell: Holland Environment by Class Size by Course Level

		Course Level		
	Class Size	Lower	Upper	Total
Artistic	Small	451	320	771
	Large	79	35	114
	Total	530	355	885
Social	Small	85	65	150
	Large	62	35	97
	Total	147	100	247
Realistic	Small	101	113	214
	Large	36	87	123
	Total	137	200	337
Enterprising	Small	164	253	417
	Large	94	158	252
	Total	258	411	669
Investigative	Small	925	476	1,401
	Large	703	357	1,060
	Total	1,628	833	2,461
All Environments	Small	1,726	1,227	2,953
	Large	974	672	1,646
	Total	2,700	1,899	4,599

designs). Specifically, ANOVA behaves in unpredictable ways when unequal cell sizes accompany non-normal distributions (Wilcox, 1998). To address this problem, traditional ANOVA procedures were followed by post-hoc analyses using robust techniques, such as those described by Wilcox (2005). This included using bootstrapping to more accurately estimate standard errors and confidence intervals while making fewer assumptions about the shape of the sampling distribution. Traditional ANOVA assumes that the sampling distribution for the mean is normal. Bootstrap ANOVA does not rely on this assumption – rather, it provides a way of estimating the properties of a sampling distribution from the data (Efron, 1979; Efron & Tibshirani, 1993b).

Knowing that key assumptions in ANOVA were likely to be violated, I used bootstrap techniques to draw conclusions about the research questions posed. For research questions one through six[5], the data were treated as a population from which 10,000 resamples – called bootstrap samples – were drawn. The size of each bootstrap sample was the same as the *N* (= 4,599) for the original data set. Bootstrapping refers to resampling *with replacement* - that means that a rating was drawn at random from the original data, added to the bootstrap sample, and replaced in the original data set before the next rating was drawn. Because the ratings were replaced in the original data set, one rating could be added to a bootstrap sample multiple times while another rating might never be added. The mean of each bootstrap sample was calculated and saved, and the resulting 10,000 means were used to estimate the sampling distribution. The standard error was estimated from the standard deviation of the sampling distribution created from the bootstrap samples.

The bias-corrected and accelerated (BCa) method was used to compute bootstrap

[5] For research questions one through six, the analyses were done using IBM SPSS (Version 21).

confidence intervals on the mean for the first six research questions. If the mean of the bootstrap distribution does not equal the sample mean, the BCa method helps to correct the bias. BCa bootstrap confidence intervals tend to be more accurate than intervals calculated using the percentile method, although they are more computationally intensive (Efron & Tibshirani, 1993a).

Although bootstrapping is most often used to estimate the accuracy of a point estimate such as a sample mean, the logic of bootstrapping can be applied to a variety of statistics (Wright, London, & Field, 2011). As such, for research question seven[6], bootstrapping was used to calculate probability estimates in a situation for which standard statistical tests have not been developed. This research question involved comparing the eta squared from the ANOVA for academic college with the eta squared from the ANOVA for Holland environment. To determine the statistical significance of a difference between the two eta squared values (η^2 for college - η^2 for environment), a bootstrap procedure was used to estimate a sampling distribution for the difference.

As in the previous analyses, the original data was treated as a population from which 10,000 bootstrap samples containing N (= 4,599) values, were created. The resampling was done within the cells of a two-way table that crossed the two classification methods. Table 5 shows the number of class ratings in each cell of this two-way table. Each bootstrap sample replicated this table in terms of the number of ratings per cell. For example, a rating drawn from the College of Science and realistic categories in the original data would be added to the College of Science and realistic categories in the bootstrap sample. Eta squared was then calculated for each

[6] For research question seven, the analyses were done using Resampling Stats for Excel (Version 4.0), an add-in for Microsoft Excel that facilitates bootstrapping and permutation procedures.

Table 5

Crosstab of Six Academic Colleges by Five Holland Environments: Number of Class Ratings in Each Category

	Holland Environment					
Academic College	Artistic	Social	Realistic	Enterprising	Investigative	Total
Liberal Arts & Human Sciences	734	146	56	397	89	1,422
Agriculture & Life Sciences	0	101	97	0	154	352
Architecture	151	0	27	26	0	204
Business	0	0	0	246	180	426
Science	0	0	0	0	1,394	1,394
Engineering	0	0	157	0	644	801
Total	885	247	337	669	2,461	4,599

classification method by conducting one-way ANOVAs on the resampled data and dividing the between-group Sums of Squares by the total Sums of Squares. At the end of each iteration, the difference between the eta squared values was saved. The result was a list of 10,000 values for the eta squared difference. The distribution of these values provided an estimate of the sampling distribution. By sorting the values and finding the 2.5^{th} and 97.5^{th} percentiles, the bootstrap 95% confidence interval for the eta squared difference was estimated. Additional details about the bootstrap procedure are provided in Appendix D.

Chapter Four

Results

The data for this study were analyzed to examine the relationship between course characteristics and student ratings of instruction. Specifically, I investigated to what extent the academic field, level, and size of a class are associated with variation in mean class ratings. Two different approaches to categorizing courses into academic fields were compared - one based on the University's academic college system and one based on Holland's (1997) theory of academic environments. The analysis began with an examination of the main effects for each of the variables of interest. Interactions among the variables were examined next. Last, the two different approaches to categorizing courses into academic fields were compared. Bootstrapping was used to calculate robust estimations of standard errors and confidence intervals for all post hoc tests.

Descriptives

Descriptive statistics for the Academic College x Class Size x Course Level design are shown in Table 6. Descriptive statistics for the Holland Environment x Class Size x Course Level design are shown in Table 7.

Academic College

The first research question asked whether differences in ratings are associated with the University's academic colleges. A one-way ANOVA was used to test differences between the college means in the population. Results were statistically significant, $F(5, 4593) = 66.05$, $p < .001$, $\eta^2 = .067$, indicating that mean ratings differ across colleges[7]. Cohen's (1988) rules of

[7] Eta squared (η^2), a measure of effect size, is calculated as the between-group Sum of Squares divided by the total Sum of Squares.

Table 6

Means, (Standard Deviations), Cell Sizes, and Ranges: Academic College by Class Size by Course Level

		Course Level		
	Class Size	Lower	Upper	Overall
Liberal Arts & Human Sciences	Small			
	n	680	440	1,120
	Mean (SD)	5.21 (.59)	5.33 (.53)	5.25 (.57)
	Range	2.10 – 6.00	3.25 – 6.00	2.10 – 6.00
	Large			
	n	222	80	302
	Mean (SD)	5.18 (.49)	5.20 (.60)	5.18 (.52)
	Range	2.86 – 5.94	3.00 – 5.94	2.86 – 5.94
	Overall			
	n	902	520	1,422
	Mean (SD)	5.20 (.57)	5.31 (.54)	5.24 (.56)
	Range	2.10 – 6.00	3.00 – 6.00	2.10 – 6.00
Agriculture & Life Sciences	Small			
	n	85	151	236
	Mean (SD)	5.34 (.44)	5.25 (.57)	5.28 (.53)
	Range	4.00 – 6.00	2.90 – 6.00	2.90 – 6.00
	Large			
	n	59	57	116
	Mean (SD)	5.08 (.78)	5.15 (.48)	5.12 (.65)
	Range	2.36 – 5.96	3.81 – 5.83	2.36 – 5.96
	Overall			
	n	144	208	352
	Mean (SD)	5.23 (.61)	5.22 (.55)	5.23 (.57)
	Range	2.36 – 6.00	2.90 – 6.00	2.36 – 6.00
Architecture	Small			
	n	56	96	152
	Mean (SD)	5.26 (.55)	5.21 (.66)	5.23 (.62)
	Range	3.47 – 6.00	3.06 – 6.00	3.06 – 6.00
	Large			
	n	2	31	52

	Mean (SD)	4.91 (.74)	4.82 (.67)	4.86 (.70)
	Range	2.73 – 5.78	3.38 – 5.81	2.73 – 5.81
	Overall			
	n	77	127	204
	Mean (SD)	5.17 (.62)	5.12 (.68)	5.14 (.66)
	Range	2.73 – 6.00	3.06 – 6.00	2.73 – 6.00
Business	Small			
	n	27	163	190
	Mean (SD)	5.32 (.53)	4.95 (.75)	5.00 (.73)
	Range	3.87 – 5.88	1.78 – 6.00	1.78 – 6.00
	Large			
	n	73	163	236
	Mean (SD)	4.98 (.63)	4.99 (.70)	4.99 (.68)
	Range	2.30 – 5.86	2.53 – 6.00	2.30 – 6.00
	Overall			
	n	100	326	426
	Mean (SD)	5.07 (.62)	4.97 (.72)	4.99 (.70)
	Range	2.30 – 5.88	6.00 – 4.97	1.78 – 6.00
Science	Small			
	n	658	172	830
	Mean (SD)	4.86 (.73)	5.11 (.62)	4.91 (.72)
	Range	1.83 – 6.00	2.30 – 6.00	1.83 – 6.00
	Large			
	n	462	102	564
	Mean (SD)	4.79 (.74)	4.74 (.86)	4.78 (.76)
	Range	1.76 – 5.88	1.86 – 5.97	1.76 – 5.97
	Overall			
	n	1120	274	1,394
	Mean (SD)	4.83 (.74)	4.97 (.74)	4.86 (.74)
	Range	1.76 – 6.00	1.86 – 6.00	1.76 – 6.00
Engineering	Small			
	n	220	205	425
	Mean (SD)	4.89 (.69)	4.89 (.67)	4.89 (.68)
	Range	2.73 – 6.00	2.70 – 6.00	2.70 -6.00
	Large			

	n	137	239	376
	Mean (SD)	4.80 (.76)	4.82 (.74)	4.81 (.74)
	Range	2.14 – 5.96	2.63 – 5.91	2.14 – 5.96
	Overall			
	n	357	444	801
	Mean (SD)	4.85 (.72)	4.85 (.71)	4.85 (.71)
	Range	2.14 – 6.00	2.63 – 6.00	2.14 – 6.00
All Colleges	Small			
	n	1,726	1,227	2,953
	Mean (SD)	5.04 (.68)	5.15 (.63)	5.09 (.66)
	Range	1.83 – 6.00	1.78 – 6.00	1.78 – 6.00
	Large			
	n	974	672	1,646
	Mean (SD)	4.92 (.71)	4.92 (.73)	4.92 (.71)
	Range	1.76 – 5.95	1.86 – 6.00	1.76 – 6.00
	Overall			
	n	2,700	1,899	4,599
	Mean (SD)	5.00 (.69)	5.07 (.68)	5.03 (.69)
	Range	1.76 – 6.00	1.78 – 6.00	1.76 – 6.00

Note. Overall = the sample size, mean, and range of all ratings in a given category.

Table 7

Means, (Standard Deviations), Cell Sizes, and Ranges: Holland Environment by Class Size by Course Level

		Course Level		
	Class Size	Lower	Upper	Overall
Artistic	Small			
	n	451	320	771
	Mean (SD)	5.28 (.54)	5.38 (.51)	5.32 (.53)
	Range	2.10 – 6.00	3.06 – 6.00	2.10 – 6.00
	Large			
	n	79	35	114
	Mean (SD)	5.19 (.46)	5.21 (.57)	5.19 (.50)
	Range	4.01 – 5.89	3.77 – 5.86	3.77 – 5.89
	Overall			
	n	530	355	885
	Mean (SD)	5.26 (.53)	5.36 (.52)	5.30 (.53)
	Range	2.10 – 6.00	3.06 – 6.00	2.10 – 6.00
Social	Small			
	n	85	65	150
	Mean (SD)	5.07 (.66)	5.22 (.65)	5.13 (.66)
	Range	2.58 – 6.00	2.90 – 6.00	2.58 – 6.00
	Large			
	n	62	35	97
	Mean (SD)	5.08 (.68)	5.12 (.60)	5.09 (.65)
	Range	2.36 – 5.92	3.57 – 5.94	2.36 – 5.94
	Overall			
	n	147	100	247
	Mean (SD)	5.07 (.66)	5.18 (.63)	5.12 (.65)
	Range	2.36 – 6.00	2.90 – 6.00	2.36 – 6.00
Realistic	Small			
	n	101	113	214
	Mean (SD)	5.29 (.60)	5.12 (.60)	5.20 (.61)
	Range	3.31 – 6.00	3.45 – 6.00	3.31 – 6.00
	Large			
	n	36	87	123

	Mean (SD)	5.21 (.68)	4.81 (.79)	4.93 (.78)
	Range	2.40 – 5.96	2.63 – 5.91	2.40 – 5.96
	Overall			
	n	137	200	337
	Mean (SD)	5.27 (.62)	4.99 (.71)	5.10 (.69)
	Range	2.40 – 6.00	2.63 – 6.00	2.40 – 6.00
Enterprising	Small			
	n	164	253	417
	Mean (SD)	5.06 (.62)	5.14 (.61)	5.11 (.62)
	Range	2.95 – 5.91	3.20 – 6.00	2.95 – 6.00
	Large			
	n	94	158	252
	Mean (SD)	5.12 (.53)	5.00 (.68)	5.05 (.63)
	Range	2.73 – 5.94	2.65 – 6.00	2.65 – 6.00
	Overall			
	n	258	411	669
	Mean (SD)	5.08 (.59)	5.09 (.64)	5.08 (.62)
	Range	2.73 – 5.94	2.65 – 6.00	2.65 – 6.00
Investigative	Small			
	n	925	476	1,401
	Mean (SD)	4.89 (.72)	5.01 (.69)	4.93 (.71)
	Range	1.83 – 6.00	1.78 – 6.00	1.78 – 6.00
	Large			
	n	703	357	1,060
	Mean (SD)	4.83 (.73)	4.86 (.74)	4.84 (.74)
	Range	1.76 – 5.96	1.86 –5.97	1.76 – 5.97
	Overall			
	n	1,628	833	2,461
	Mean (SD)	4.87 (.72)	4.95 (.71)	4.89 (.72)
	Range	1.76 – 6.00	1.78 – 6.00	1.76 – 6.00
All Environments	Small			
	n	1,726	1,227	2,953
	Mean (SD)	5.04 (.68)	5.15 (.63)	5.09 (.66)
	Range	1.83 – 6.00	1.78 – 6.00	1.78 – 6.00
	Large			

n	974	672	1,646
Mean (SD)	4.92 (.71)	4.92 (.73)	4.92 (.71)
Range	1.76 – 5.95	1.86 – 6.00	1.76 – 6.00
Overall			
n	2,700	1,899	4,599
Mean (SD)	5.00 (.69)	5.07 (.68)	5.03 (.69)
Range	1.76 – 6.00	1.78 – 6.00	1.76 – 6.00

Note. Overall = the sample size, mean, and range of all ratings in a given category.

thumb for eta squared (η^2), classify this as a medium effect[8].

The Games-Howell[9] procedure was used to make pairwise comparisons between all of the means. Each college mean differed significantly from at least one other college mean. Table 8 shows the significant differences found between college means, and the bootstrapped standard errors and confidence intervals. Figure 2 is a graphic depiction of the means in the six colleges.

Levene's test[10] was used to assess the homogeneity of variance. Results were significant, $F(5, 4593) = 25.41$, $p < .001$, indicating that the variance of ratings differs across colleges. Table 9 shows that within group variance ranges from a high of .54 in the College of Science to a low of .31 in the College of Liberal Arts and Human Sciences. Figure 3 plots variance against the mean for ratings in each college; it shows a tendency for within group variance to become smaller as college means become higher.

Holland Environment

The second research question asked if differences in ratings are associated with Holland's academic environments. A one-way ANOVA was used to test differences between the environment means in the population. Results were statistically significant, $F(4, 4594) = 66.22$, $p < .001$, $\eta^2 = .055$, indicating that mean ratings differ across Holland environments. Cohen's (1988) rules of thumb for η^2 classify this as a small effect. The Games-Howell procedure was used to make pairwise comparisons between all of the means. Each environment mean was

---

[8] Cohen's (1988, pp. 285 - 287) rule of thumb for effect sizes: $\eta^2 = .010$ is small, $\eta^2 = .059$ is medium, $\eta^2 = .138$ is large.

[9] The Games-Howell test was used for all pairwise multiple comparisons because it maintains the family-wise Type I error rate at the desired level in situations where cell sizes and/or variances are unequal (Games & Howell, 1976).

[10] Levene's (1960) test performs a one-way ANOVA on the deviation scores to determine whether variances in the different groups are equal.

Table 8

Comparison of Mean Ratings, Standard Errors, and Confidence Intervals in Six Academic Colleges

College	*n*	Mean (*SE*)	95% CI[a]	Comparison of Means[b]
(1) Liberal Arts & Human Sciences	1,422	5.24 (.02)	[5.21, 5.27]	**1** > 4, 5, 6
(2) Agriculture & Life Sciences	352	5.23 (.03)	[5.16, 5.29]	**2** > 4, 5, 6
(3) Architecture	204	5.14 (.05)	[5.04, 5.23]	**3** > 5, 6
(4) Business	426	4.99 (.03)	[4.93, 5.06]	1, 2 > **4** > 5, 6
(5) Science	1,394	4.86 (.02)	[4.82, 4.90]	1, 2, 3, 4 > **5**
(6) Engineering	801	4.85 (.03)	[4.80, 4.90]	1, 2, 3, 4, > **6**
Marginal	4,599	5.05 (.01)	[5.03, 5.07]	

Note. Numbers in parentheses next to college names refer to the numbers used for illustrating differences in the "comparisons" column. Standard errors and confidence intervals are based on 10,000 bootstrap samples.

[a]CI = bias corrected and accelerated confidence interval for the mean. [b]Differences are significant with family-wise $\alpha = .05$.

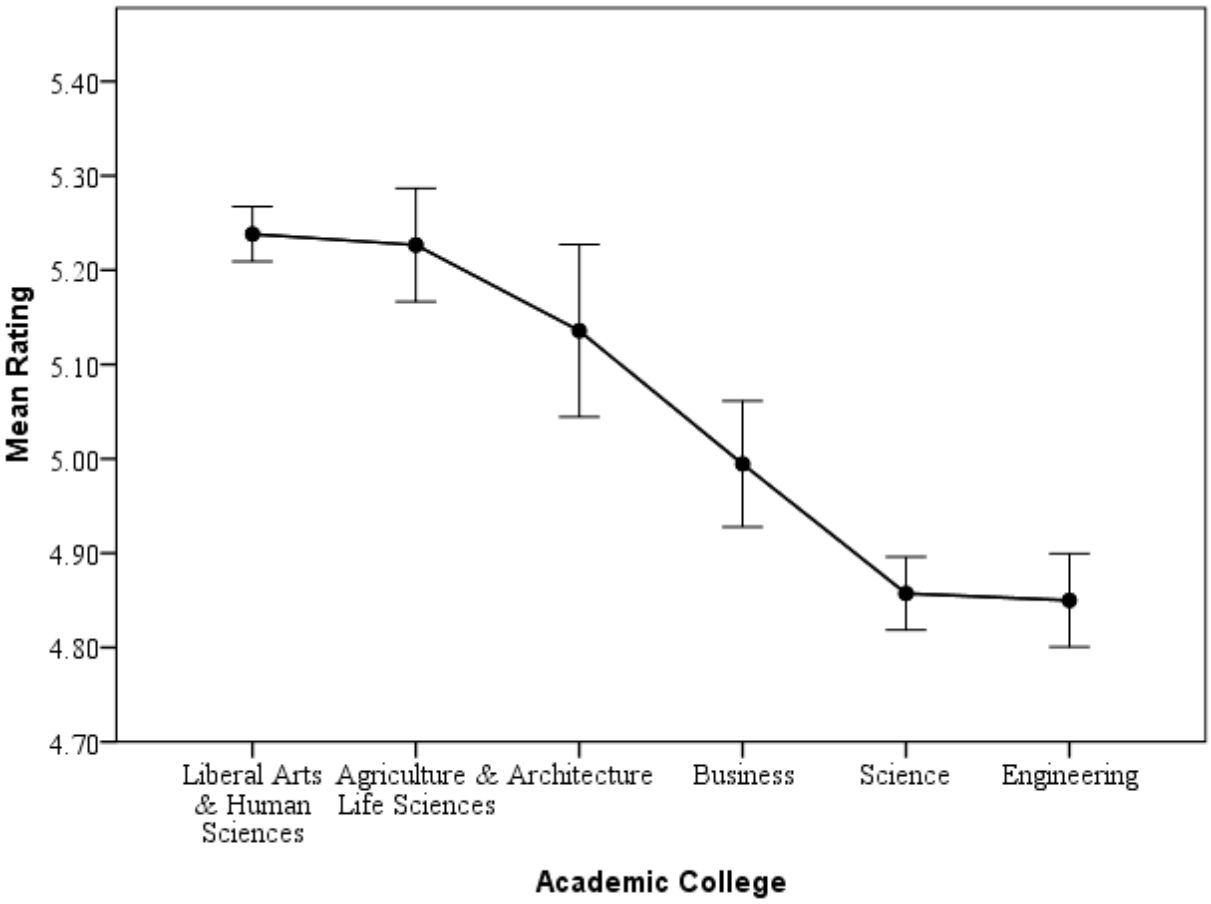

Figure 2. Plot of Means and 95% Confidence Intervals for Six Academic Colleges

Table 9

Indicators of the Distribution of Ratings in Six Academic Colleges

College	*n*	Mean	Variance	*SD*	Skew	Kurtosis
Liberal Arts & Human Sciences	1,422	5.24	.31	.56	-1.40	2.61
Agriculture & Life Sciences	352	5.23	.33	.57	-1.58	3.43
Architecture	204	5.14	.44	.66	-1.19	1.15
Business	426	4.99	.49	.70	-1.42	2.32
Science	1,394	4.86	.54	.74	-1.10	1.28
Engineering	801	4.85	.51	.71	-.74	.24

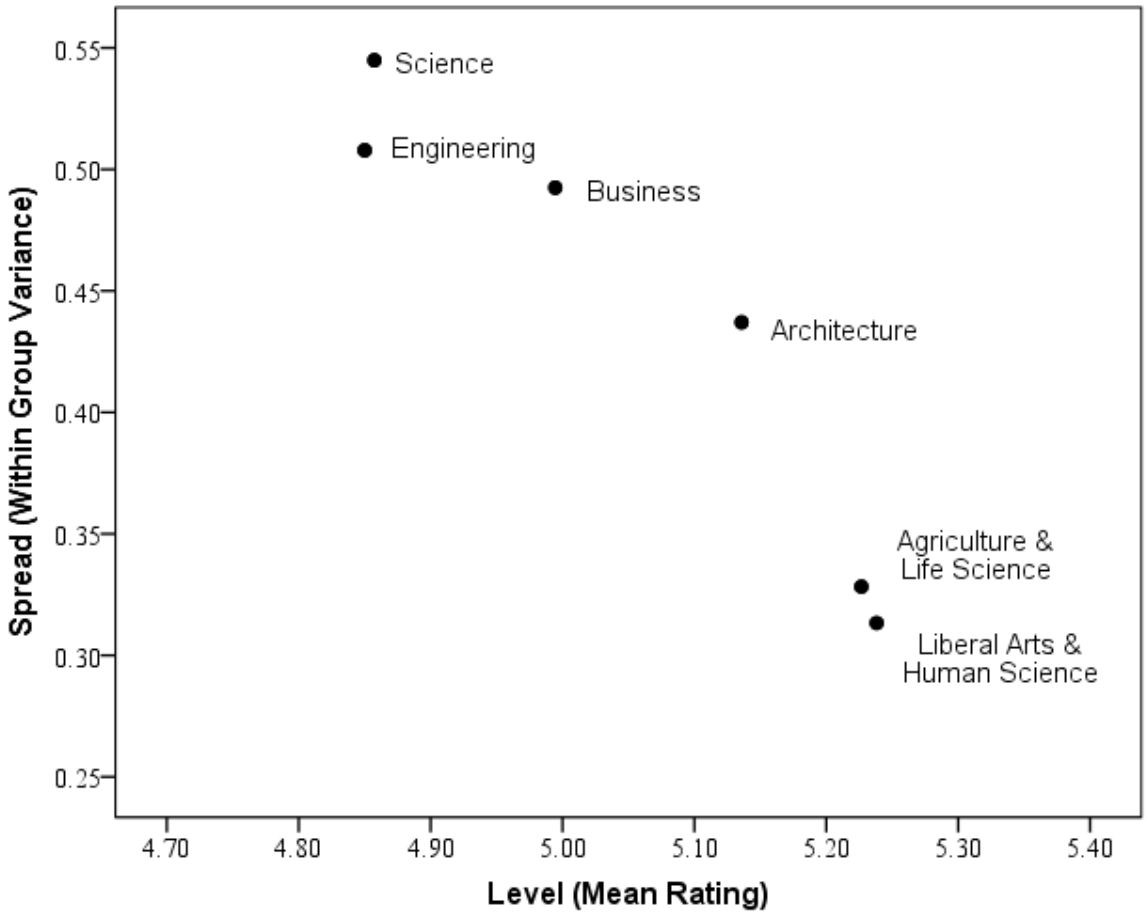

Figure 3. Plot of Spread versus Level for Ratings in Six Academic Colleges

found to be significantly different from at least one other environment mean. Table 10 shows the significant differences found between environment means, and the bootstrapped standard errors and confidence intervals. Figure 4 shows the graphic depiction of the means.

Levene's test was used to assess the homogeneity of variance. Results were significant, $F(4, 4594) = 26.87$, $p < .001$, indicating that the variance of ratings differs across Holland environments. Table 11 shows that within group variance ranges from a high of .52 in the investigative environment to a low of .28 in the artistic environment. Figure 5 plots variance against the mean for ratings in each environment; it shows that the smallest variance accompanies the highest environment mean while the largest variance accompanies the lowest environment mean. Relatively moderate variance accompanies means that are clustered in the center of the distribution.

Course Level

The third research question asked if differences in ratings are associated with course level. A one-way ANOVA was used to test the difference between the two course level means in the population. Results were statistically significant, $F(1, 4597) = 13.45$, $p < .001$, $\eta^2 = .003$, indicating that upper level classes tend to receive higher ratings than lower level classes. However, Cohen's (1988) rules of thumb for η^2 classify this as a trivial effect, indicating that the difference may not be a meaningful one. Table 12 shows mean ratings for the two course levels, and the bootstrapped standard errors and confidence intervals.

Levene's test was used to assess the homogeneity of variance. Results indicated a non-significant difference in the variance of ratings across course levels, $F(1, 4597) = .80$, $p = .372$. Table 13 shows variances of .48 in the lower course level and .46 in the upper course level.

Table 10

Comparison of Mean Ratings, Standard Errors, and Confidence Intervals in Five Holland Environments

Environment	*n*	Mean (*SE*)	95% CI[a]	Comparison of Means[b]
(1) Artistic	885	5.30 (.02)	[5.27, 5.34]	**1** > 2, 3, 4, 5
(2) Social	247	5.12 (.04)	[5.03, 5.19]	1 > **2** > 5
(3) Realistic	337	5.10 (.04)	[5.03, 5.17]	1 > **3** > 5
(4) Enterprising	669	5.08 (.02)	[5.04, 5.13]	1> **4** > 5
(5) Investigating	2,461	4.89 (.01)	[4.86, 4.92]	1, 2, 3, 4 > **5**
Marginal	4,599	5.10 (.01)	[5.07, 5.13]	

Note. Numbers in parentheses next to Holland environment names refer to the numbers used for illustrating differences in the "comparison of means" column. Standard errors and confidence intervals are based on 10,000 bootstrap samples.

[a]CI = bias corrected and accelerated confidence interval for the mean. [b]Differences are significant with family-wise $\alpha = .05$.

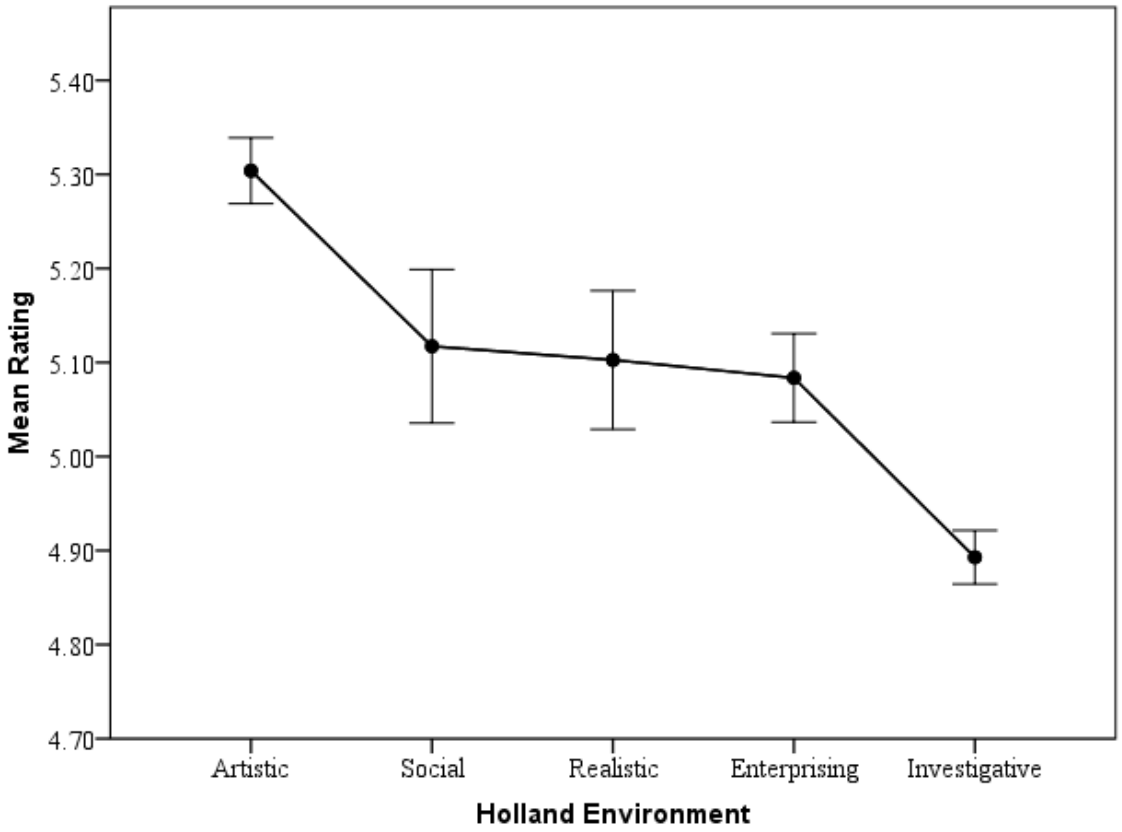

Figure 4. Plot of Means and 95% Confidence Intervals for Five Holland Environments

Table 11

Indicators of the Distribution of Ratings in Five Holland Environments

Environment	*n*	Mean	Variance	*SD*	Skew	Kurtosis
Artistic	885	5.30	.28	.53	-1.45	3.03
Social	247	5.12	.42	.65	-1.55	3.11
Realistic	337	5.10	.47	.69	-1.25	1.45
Enterprising	669	5.08	.39	.62	-1.17	1.23
Investigative	2,461	4.89	.52	.72	-1.08	1.20

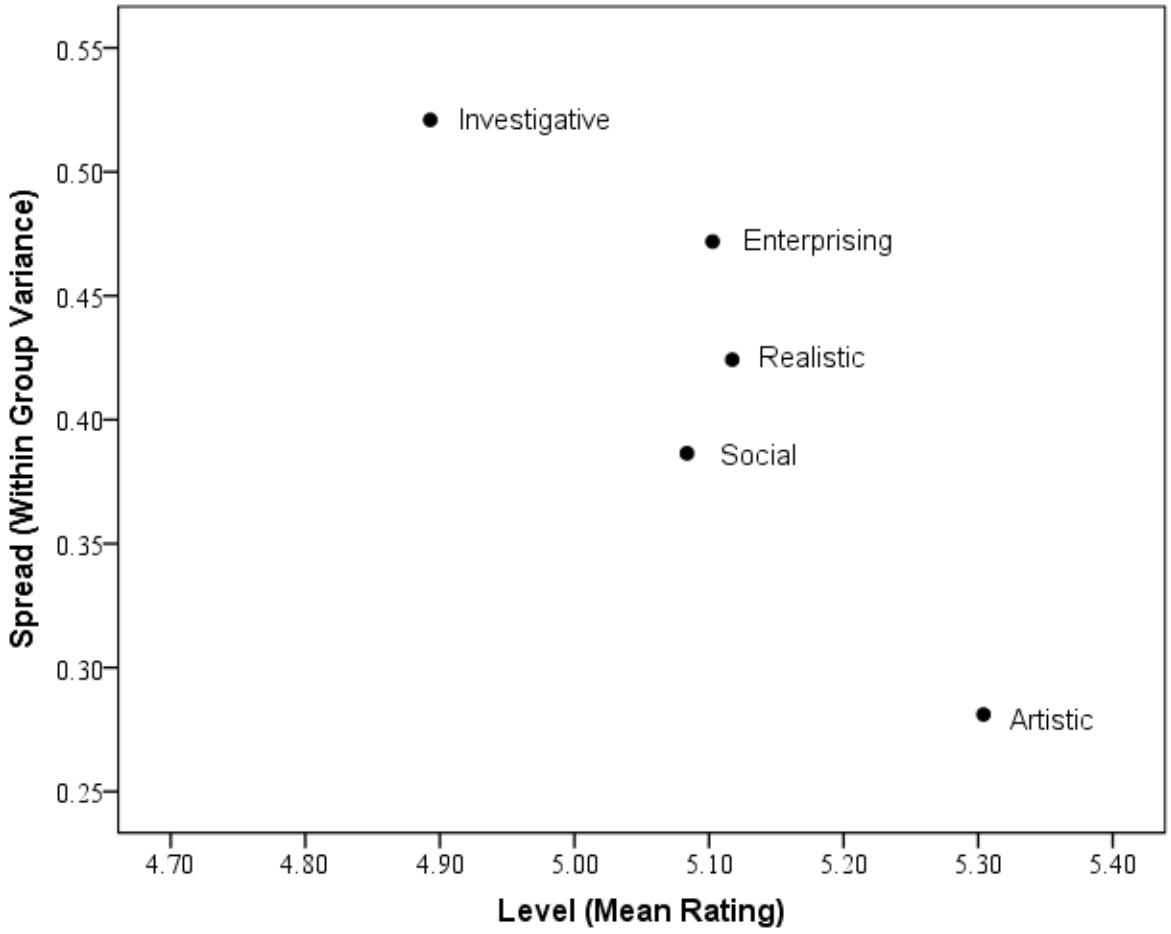

Figure 5. Plot of Spread versus Level for Ratings in Five Holland Environments

Table 12

Mean Ratings, Standard Errors, and Confidence Intervals in Two Course Levels

Course Level	*n*	Mean (*SE*)	95% CI[a]
Upper	1,899	5.07 (.02)	[5.04, 5.10]
Lower	2,700	5.00 (.01)	[4.97, 5.02]
Marginal	4,599	5.03 (.01)	[5.01, 5.05]

Note. Standard errors and confidence intervals are based on 10,000 bootstrap samples.

[a]CI = bias corrected and accelerated confidence interval for the mean.

Table 13

Indicators of the Distribution of Ratings in Two Course Levels

Course Level	*n*	Mean	Variance	*SD*	Skew	Kurtosis
Upper	1,899	5.07	.46	.68	-1.24	1.62
Lower	2,700	5.00	.48	.69	-1.19	1.61

Class Size

The fourth research question asked if differences in ratings are associated with class size. A one-way ANOVA indicated that the difference between the two class size means in the population was statistically significant, F(1, 4597) = 66.58, $p < .001$, $\eta^2 = .014$, with small classes receiving higher ratings than large classes. Cohen's (1988) guidelines for η^2 classify this as a small (still meaningful) effect size. Table 14 shows mean ratings for the two class sizes, and the bootstrapped standard errors and confidence intervals.

Levene's test was used to assess the homogeneity of variance. Results were significant, $F(1, 4597) = 8.15$, $p = .004$, indicating that the variance of ratings differs across class sizes. Table 15 shows variances of .51 in the large class size and .44 in the small class size.

Academic College x Course Level x Class Size

The fifth research question asked if there are interaction effects between academic college and course level or class size. A 6 x 2 x 2 ANOVA was used to analyze the group means using Type III Sums of Squares. There was a significant main effect of college, $F(5, 4575) = 38.95$, $p < .001$, $\eta^2 = .039$, a significant main effect of class size, $F(1, 4575) = 37.14$, $p < .001$, $\eta^2 = .007$, and a non-significant main effect of course level, $F(1, 4575) = .20$, $p = .655$, $\eta^2 < .001$ However, the three-way interaction was significant, $F(5, 4575) = 3.41$, $p = .004$, $\eta^2 = .003$, indicating that the interaction effect of class size and course level is not the same in all colleges. Thus, the main effects and the two-way interaction effects should be interpreted with caution because interaction among the three variables is associated with differences between the ratings. The ANOVA summary for this analysis is shown in Table 16.

Following the finding of a significant three-way interaction, follow up tests were conducted to facilitate understanding of the interaction of class size and course level within each

Table 14

Mean Ratings, Standard Errors, and Confidence Intervals in Two Class Sizes

Class Size	*n*	Mean (*SE*)	95% CI[a]
Small	2,953	5.09 (.01)	[5.06, 5.11]
Large	1,646	4.92 (.02)	[4.88, 4.95]
Marginal	4,599	5.00 (.01)	[4.98, 5.02]

Note. Standard errors and confidence intervals are based on 10,000 bootstrap samples.

[a]CI = bias corrected and accelerated confidence interval for the mean.

Table 15

Indicators of the Distribution of Ratings in Two Class Sizes

Class Size	*n*	Mean	Variance	*SD*	Skew	Kurtosis
Small	2,953	5.09	.44	.66	-1.20	156
Large	1,646	4.92	.51	.71	-1.21	1.57

Table 16

ANOVA for Ratings: Academic College by Class Size by Course Level

Source	*df*	*Type III SS*	*MS*	*F*	*p*	η^2
College	5	84.79	16.96	38.95	< .001	.039
Class Size	1	16.17	16.17	37.14	< .001	.007
Course Level	1	.09	.09	.20	.655	< .001
College x Class Size	5	4.65	.93	2.14	.058	.002
College x Course Level	5	4.62	.92	1.12	.060	.002
Class Size x Course Level	1	.03	.03	.08	.779	< .001
College x Class Size x Course Level	5	7.42	1.48	3.41	.004	.003
Error	4,575	1,991.98	.44			
Total	4,598	2,169.33				

Note. Eta squared (η^2) for each source of variation is calculated as the between-group Sum of Squares for the given source divided by the total Sum of Squares.

college. Six 2 x 2 ANOVAs were used to analyze the interaction of class size and course level within each college. As with the three-way ANOVA, these analyses were based on Type III Sums of Squares.

In the College of Liberal Arts and Human Sciences, there was a significant main effect of class size with small classes receiving higher ratings than large classes, $F(1, 1418) = 3.72$, $p = .054$, $\eta^2 = .003$, but a non-significant main effect of course level, $F(1, 1418) = 3.22$, $p = .073$, $\eta^2 = .002$. The interaction was not significant, $F(1, 1418) = 1.42$, $p = .234$, $\eta^2 = .001$. Table 17 shows the means and confidence intervals for ratings in this college. Figure 6 shows the graphic depiction of the interaction.

In the College of Agriculture and Life Sciences, there was a significant main effect of class size, indicating that small classes receive higher ratings than large classes, $F(1, 348) = 7.38$, $p = .007$, $\eta^2 = .021$, but a non-significant main effect of course level, $F(1, 348) = .02$, $p = .898$, $\eta^2 = .000$. The interaction was not significant, $F(1, 348) = 1.41$, $p = .237$, $\eta^2 = .004$. Table 18 shows the means and confidence intervals for ratings in this college. Figure 7 shows the graphic depiction of the interaction.

In the College of Architecture and Urban Studies, there was a significant main effect of class size showing that small classes receive higher ratings than large classes, $F(1, 200) = 12.29$, $p = .001$, $\eta^2 = .058$, but a non-significant main effect of course level, $F(1, 200) = .44$, $p = .506$, $\eta^2 = .002$. The interaction was not significant, $F(1, 200) = .05$, $p = .816$, $\eta^2 < .001$. Table 19 shows the means and confidence intervals for ratings in this college. Figure 8 shows the graphic depiction of the interaction.

Table 17

Equally Weighted Means and Confidence Intervals for Ratings in the College of Liberal Arts and Human Sciences (n = 1,422): Class Size by Course Level

	Course Level		
Class Size	Lower	Upper	Marginal
Small			
Mean (*SE*)	5.21 (.02)	5.33 (.03)	5.27 (.02)
95% CI	[5.16, 5.25]	[5.28, 5.37]	[5.23, 5.30]
Large			
Mean (*SE*)	5.18 (.03)	5.20 (.07)	5.19 (.04)
95% CI	[5.11, 5.24]	[5.06, 5.33]	[5.11, 5.26]
Marginal			
Mean (*SE*)	5.19 (.02)	5.26 (.04)	5.23 (.02)
95% CI	[5.15, 5.23]	[5.19, 5.33]	[5.19, 5.27]

Note. Standard errors and confidence intervals are based on 10,000 bootstrap samples for the mean. CI = bias corrected and accelerated confidence interval for the mean.

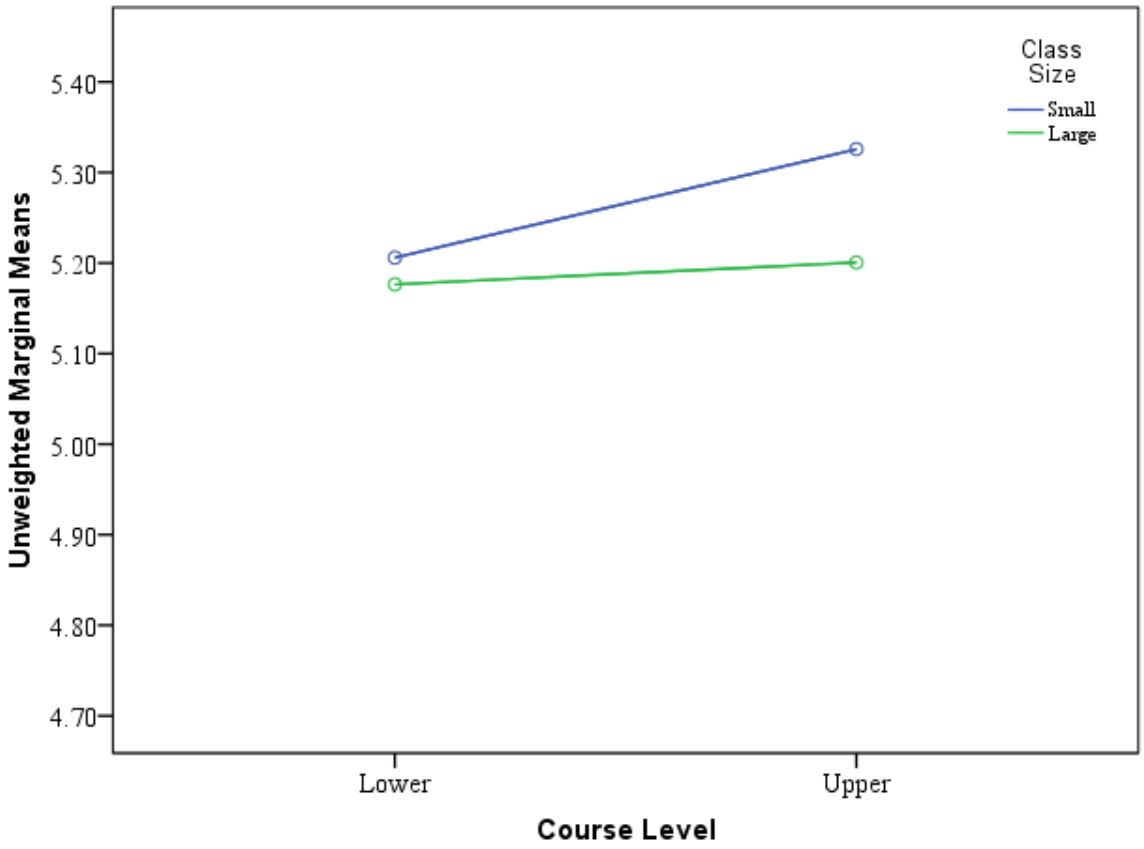

Figure 6. Plot of Equally Weighted Marginal Means for Ratings in the College of Liberal Arts and Human Sciences: Class Size by Course Level

Table 18

Equally Weighted Means and Confidence Intervals for Ratings in the College of Agriculture and Life Sciences (n=352): Class Size by Course Level

	Course Level		
Class Size	Lower	Upper	Marginal
Small			
Mean (*SE*)	5.34 (.05)	5.25 (.05)	5.29 (.03)
95% CI	[5.24, 5.43]	[5.16, 5.34]	[5.23, 5.36]
Large			
Mean (*SE*)	5.08 (.10)	5.15 (.06)	5.12 (.06)
95% CI	[4.86, 5.28]	[5.02, 5.28]	[4.99, 5.23]
Marginal			
Mean (*SE*)	5.21 (.06)	5.20 (.04)	5.23 (.03)
95% CI	[5.09, 5.32]	[5.12, 5.28]	[5.13, 5.27]

Note. Standard errors and confidence intervals are based on 10,000 bootstrap samples for the mean. CI = bias corrected and accelerated confidence interval for the mean.

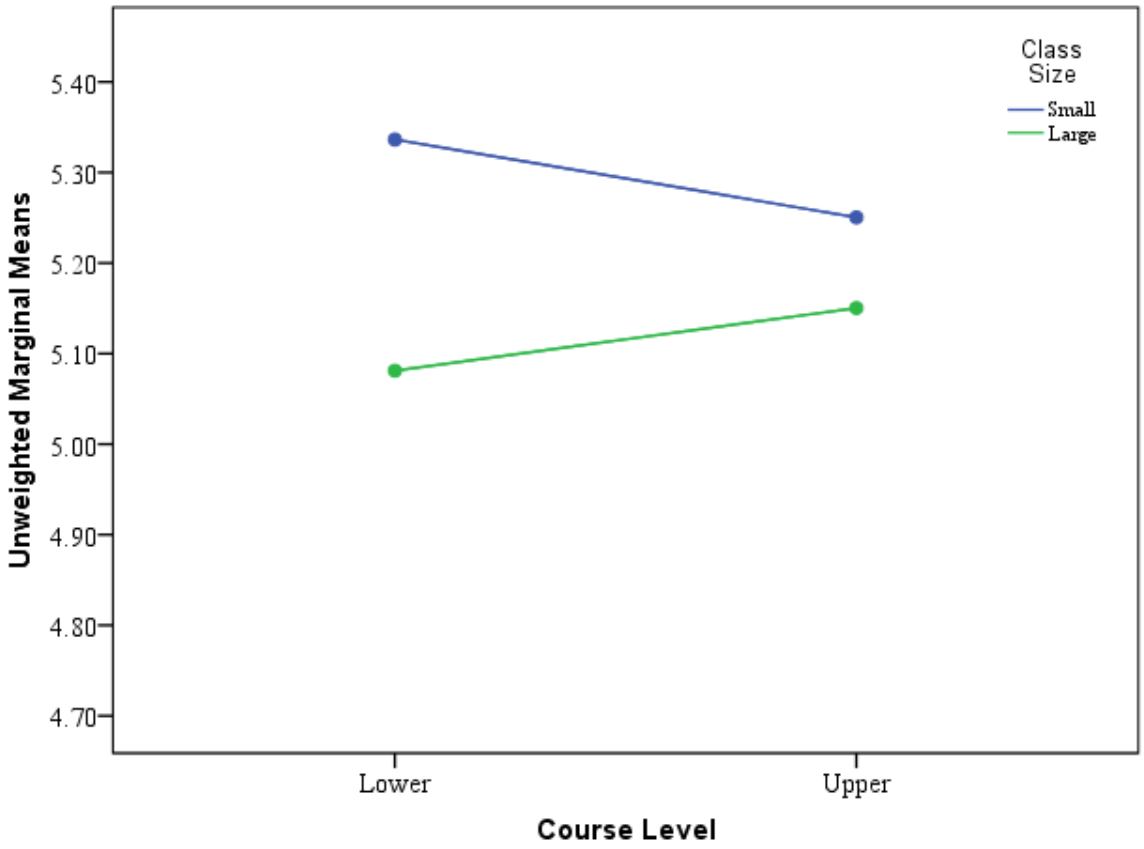

Figure 7. Plot of Equally Weighted Marginal Means for Ratings in the College of Agriculture and Life Sciences: Class Size by Course Level.

Table 19

Equally Weighted Means and Confidence Intervals for Ratings in the College of Architecture and Urban Studies (n = 204): Class Size by Course Level

Class Size	Course Level		Marginal
	Lower	Upper	
Small			
Mean (*SE*)	5.26 (.07)	5.21 (.07)	5.24 (.05)
95% CI	[5.11, 5.40]	[5.07, 5.34]	[5.14, 5.33]
Large			
Mean (*SE*)	4.91 (.16)	4.82 (.12)	4.87 (.10)
95% CI	[4.55, 5.22]	[4.59, 5.04]	[4.65, 5.06]
Marginal			
Mean (*SE*)	5.09 (.09)	5.02 (.07)	5.05 (.06)
95% CI	[4.90, 5.26]	[4.88, 5.15]	[4.94, 5.16]

Note. Standard errors and confidence intervals are based on 10,000 bootstrap samples for the mean. CI = bias corrected and accelerated confidence interval for the mean.

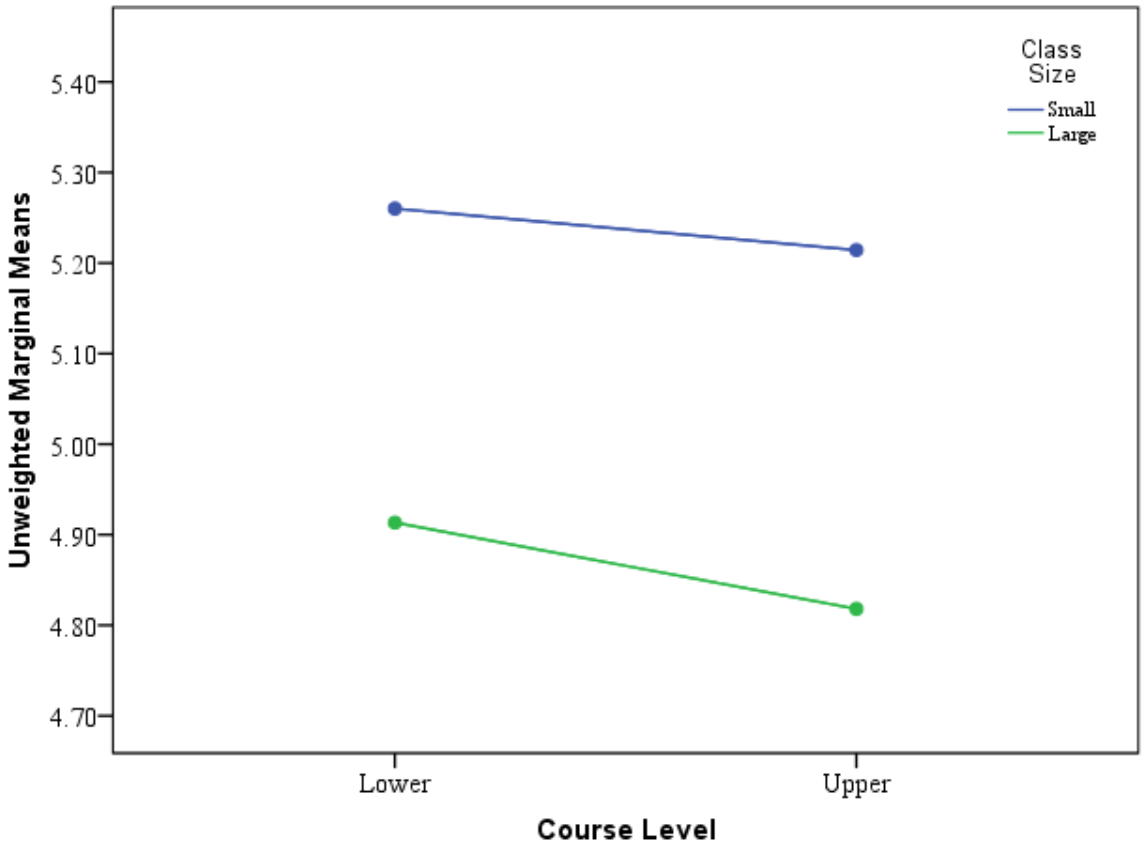

Figure 8. Plot of Equally Weighted Marginal Means for Ratings in the College of Architecture and Urban Studies: Class Size by Course Level

In the College of Business, there was a non-significant main effect of class size, $F(1, 422) = 2.88$, $p = .090$, $\eta^2 = .007$, but a significant main effect of course level, $F(1, 422) = 4.21$, $p = .041$, $\eta^2 = .010$. However, the interaction was significant, $F(1, 422) = 4.65$, $p = .032$, $\eta^2 = .011$, indicating that the relationship between course level and ratings is moderated by class size. Table 20 shows the means and confidence intervals for ratings in this college. Figure 9 shows the graphic depiction of the interaction.

In the College of Science, there was a significant main effect of class size showing that small classes receive higher ratings than large classes, $F(1, 1390) = 17.99$, $p < .001$, $\eta^2 = .131$, but a non-significant main effect of course level, $F(1, 1390) = 3.63$, $p = .057$, $\eta^2 = .003$. However the interaction was significant, $F(1, 1390) = 9.08$, $p = .003$, $\eta^2 = .006$, indicating that the relationship between class size and ratings is moderated by course level. Table 21 shows the means and confidence intervals for ratings in this college. Figure 10 shows the graphic depiction of the interaction.

In the College of Engineering, neither the main effect of class size, $F(1, 797) = 2.54$, $p = .111$, $\eta^2 = .003$, nor the main effect of course level, $F(1, 797) = .05$, $p = .828$, $\eta^2 < .001$, were significant. The interaction was not significant, $F(1, 797) = .03$, $p = .870$, $\eta^2 < .001$. Table 22 shows the means and confidence intervals for ratings in this college. Figure 11 shows the graphic depiction of the interaction.

Holland Environment x Course Level x Class Size

The sixth research question asked if there are interaction effects between Holland's academic environments and course level or class size. A 5 x 2 x 2 ANOVA was used to analyze the group means using Type III Sums of Squares. There was a significant main effect of environment, $F(4,4579) = 381.77$, $p < .001$, $\eta^2 = .026$, although there was also a significant

Table 20

Equally Weighted Means and Confidence Intervals for Ratings in the College of Business (n = 426): Class Size by Course Level

Class Size	Course Level		Marginal
	Lower	Upper	
Small			
Mean (*SE*)	5.32 (.10)	4.95 (.06)	5.13 (.06)
95% CI	[5.09, 5.53]	[4.83, 5.07]	[5.01, 5.25]
Large			
Mean (*SE*)	4.98 (.07)	4.99 (.05)	4.99 (.05)
95% CI	[4.82, 5.12]	[4.88, 5.10]	[4.89, 5.08]
Marginal			
Mean (*SE*)	5.15 (.06)	4.97 (.04)	5.06 (.04)
95% CI	[5.02, 5.28]	[4.89, 5.05]	[4.98, 5.13]

Note. Standard errors and confidence intervals are based on 10,000 bootstrap samples for the mean. CI = bias corrected and accelerated confidence interval for the mean.

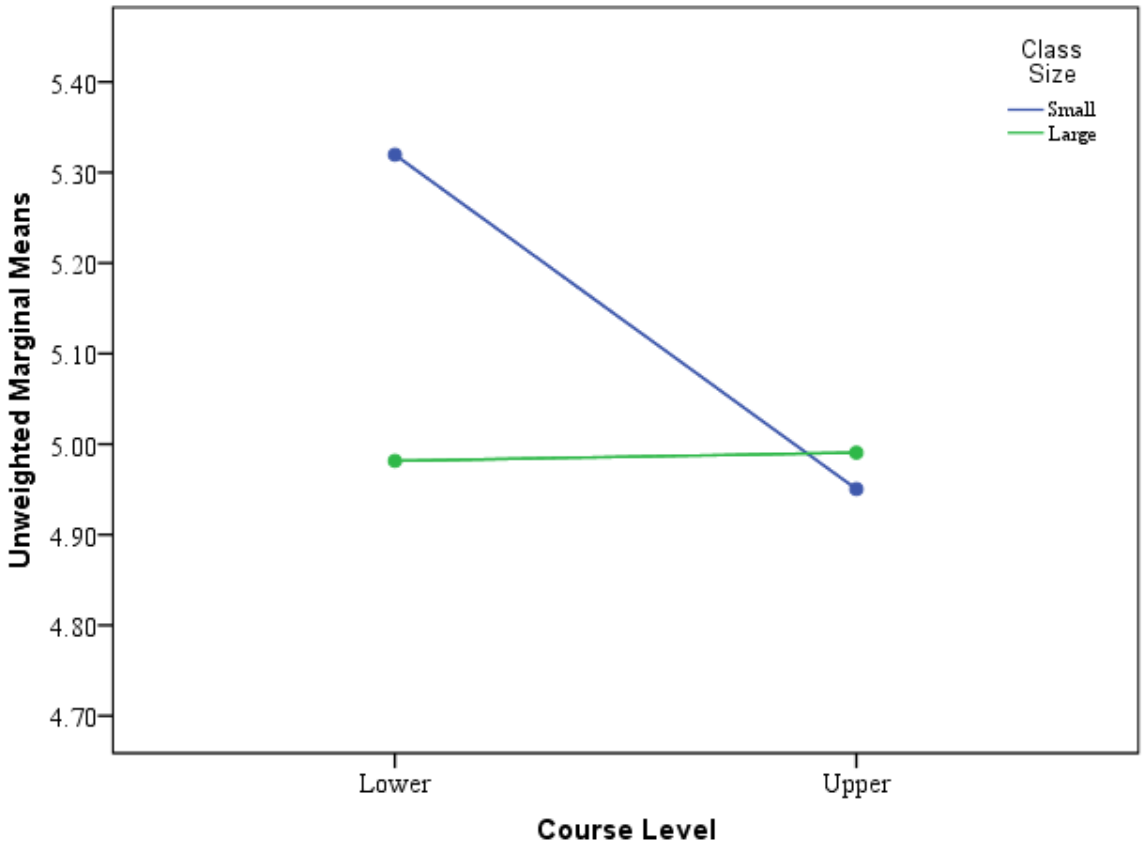

Figure 9. Plot of Equally Weighted Marginal Means for Ratings in the College of Business: Class Size by Course Level

Table 21

Equally Weighted Means and Confidence Intervals for Ratings in the College of Science (n = 1,394): Class Size by Course Level

	Course Level		
Class Size	Lower	Upper	Marginal
Small			
Mean (*SE*)	4.86 (.03)	5.11 (.05)	4.98 (.03)
95% CI	[4.80, 4.91]	[5.01, 5.20]	[4.93, 5.04]
Large			
Mean (*SE*)	4.79 (.03)	4.74 (.08)	4.77 (.05)
95% CI	[4.73, 4.86]	[4.56, 4.90]	[4.67, 4.85]
Marginal			
Mean (*SE*)	4.82 (.02)	4.92 (.05)	4.87 (.03)
95% CI	[4.78, 4.87]	[4.82, 5.01]	[4.82, 4.93]

Note. Standard errors and confidence intervals are based on 10,000 bootstrap samples for the mean. CI = bias corrected and accelerated confidence interval for the mean.

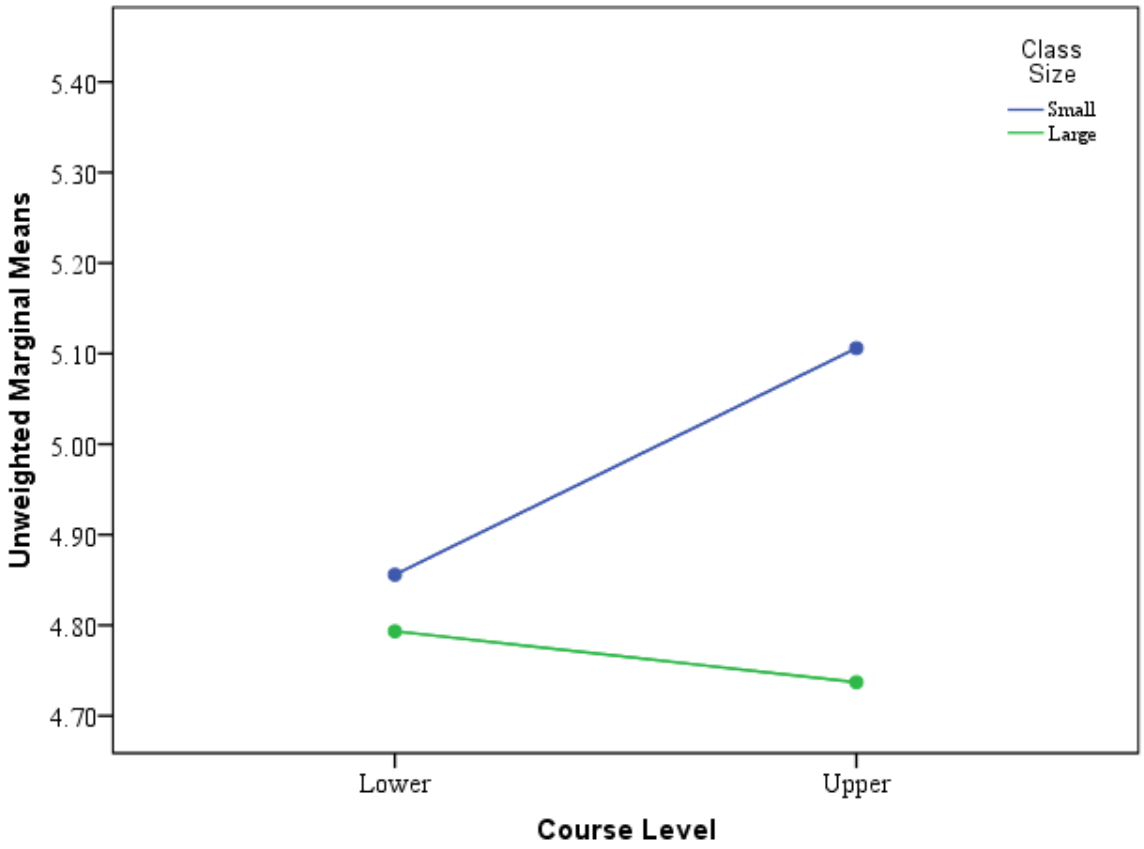

Figure 10. Plot of Equally Weighted Marginal Means for Ratings in the College of Science: Class Size by Course Level

Table 22

Equally Weighted Means and Confidence Intervals for Ratings in the College of Engineering (n = 801): Class Size by Course Level

Class Size	Course Level		Marginal
	Lower	Upper	
Small			
Mean (*SE*)	4.89 (.05)	4.89 (.05)	4.89 (.03)
95% CI	[4.79, 4.98]	[4.79, 4.98]	[4.82, 4.95]
Large			
Mean (*SE*)	4.80 (.06)	4.82 (.05)	4.81 (.04)
95% CI	[4.66, 4.92]	[4.72, 4.91]	[4.72, 4.89]
Marginal			
Mean (*SE*)	4.84 (.04)	4.85 (.03)	4.85 (.03)
95% CI	[4.76, 4.92]	[4.78, 4.92]	[4.79, 4.90]

Note. Standard errors and confidence intervals are based on 10,000 bootstrap samples for the mean. CI = bias corrected and accelerated confidence interval for the mean.

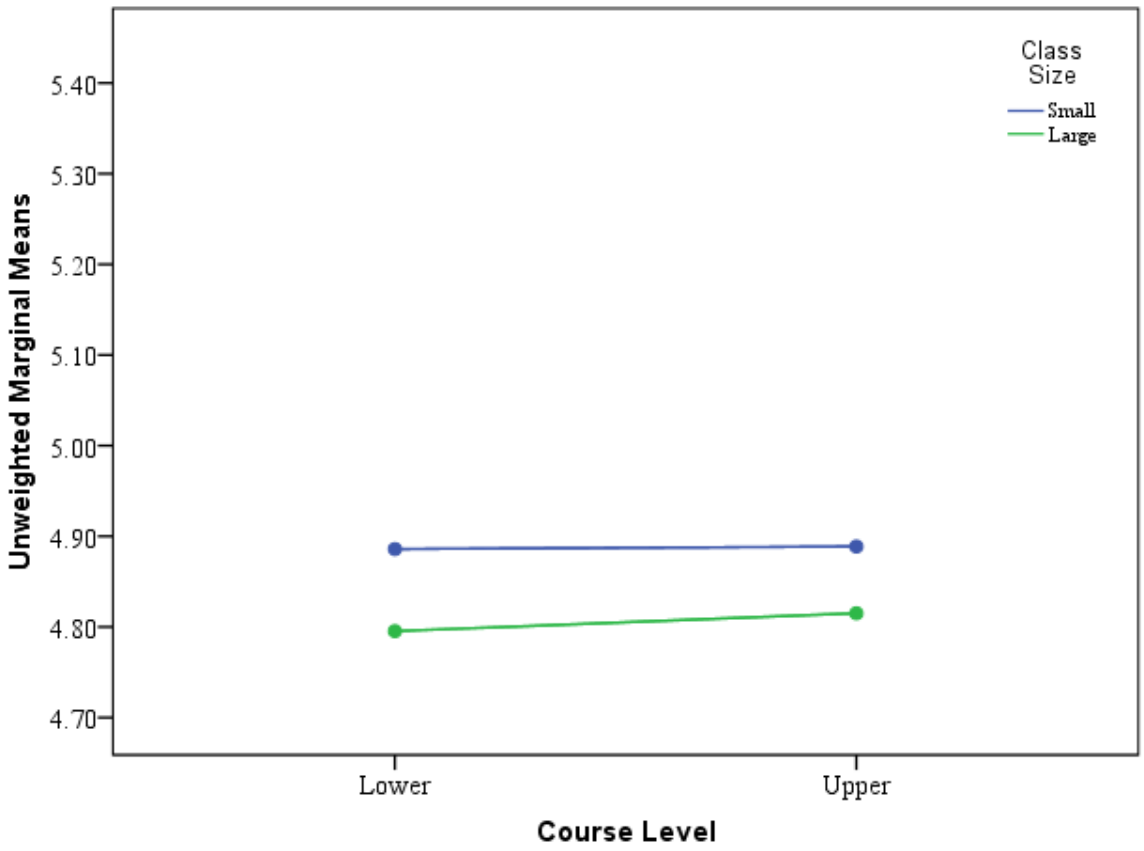

Figure 11. Plot of Equally Weighted Marginal Means for Ratings in the College of Engineering: Class Size by Course Level.

interaction between environment and course level, $F(4, 4579) = 4.90$, $p = .001$, $\eta^2 = .004$, indicating that the effect of environment is moderated by course level. There was a significant main effect of class size, $F(1,4579) = 11.27$, $p = .001$, $\eta^2 = .002$, although there was also a significant interaction between class size and course level, $F(1 ,4579) = 5.27$, $p = .022$, $\eta^2 = .001$, indicating that the effect of class size is moderated by course level. The main effect of course level was not significant, $F(1, 4579) = .22$, $p = .639$, $\eta^2 < .001$, nor was the interaction between environment and class size, $F(4, 4579) = .85$, $p = .495$, $\eta^2 = .001$. The three-way interaction was not significant, $F(4, 4579) = .43$, $p = .786$, $\eta^2 < .001$. The ANOVA summary for this analysis is shown in Table 23.

Because there are many patterns of cell means that could lead to the results found in the three-way design, follow-up tests were conducted to facilitate understanding of the interaction effects of class size and course level within each environment. Five 2 x 2 ANOVAs were used to analyze the interaction of class size and course level within each environment. As with the three-way ANOVA, these analyses were based on equally weighted means and Type III Sums of Squares.

In the artistic environment, there was a significant main effect of class size with small classes receiving higher ratings than large classes, $F(1, 881) = 5.34$, $p = .021$, $\eta^2 = .006$, but a non-significant main effect of course level, $F(1, 881) = 1.10$, $p = .295$, $\eta^2 = .001$. The interaction was not significant, $F(1, 881) = .52$, $p = .471$, $\eta^2 = .001$. Table 24 shows the means and confidence intervals for ratings in this environment. Figure 12 shows the graphic depiction of the interaction.

Table 23

ANOVA for Ratings: Holland Environment by Class Size by Course Level

Source	*df*	*Type III SS*	*MS*	*F*	*p*	η^2
Environment	4	56.14	14.04	31.77	.000	.026
Class Size	1	4.98	4.98	11.27	.001	.002
Course Level	1	.10	.10	.22	.639	< .001
Environment x Class Size	4	1.50	.37	.85	.495	.001
Environment x Course Level	4	8.66	2.17	4.90	.001	.004
Class Size x Course Level	1	2.33	2.33	5.27	.022	.001
Environment x Class Size x Course Level	4	.76	.19	.43	.786	< .001
Error	4,579	2,023.20	.44			
Total	4,598	2,169.33				

Table 24

Equally Weighted Means and Confidence Intervals for Ratings in the Artistic Environment (n = 885): Class Size by Course Level

	Course Level		
Class Size	Lower	Upper	Marginal
Small			
Mean (*SE*)	5.28 (.03)	5.38 (.03)	5.33 (.02)
95% CI	[5.23, 5.33]	[5.32, 5.43]	[5.29, 5.37]
Large			
Mean (*SE*)	5.19 (.05)	5.21 (.10)	5.20 (.05)
95% CI	[5.08, 5.29]	[5.01, 5.39]	[5.09, 5.30]
Marginal			
Mean (*SE*)	5.23 (.03)	5.29 (.05)	5.26 (.03)
95% CI	[5.17, 5.29]	[5.19, 5.39]	[5.21, 5.32]

Note. Standard errors and confidence intervals are based on 10,000 bootstrap samples for the mean. CI = bias corrected and accelerated confidence interval for the mean.

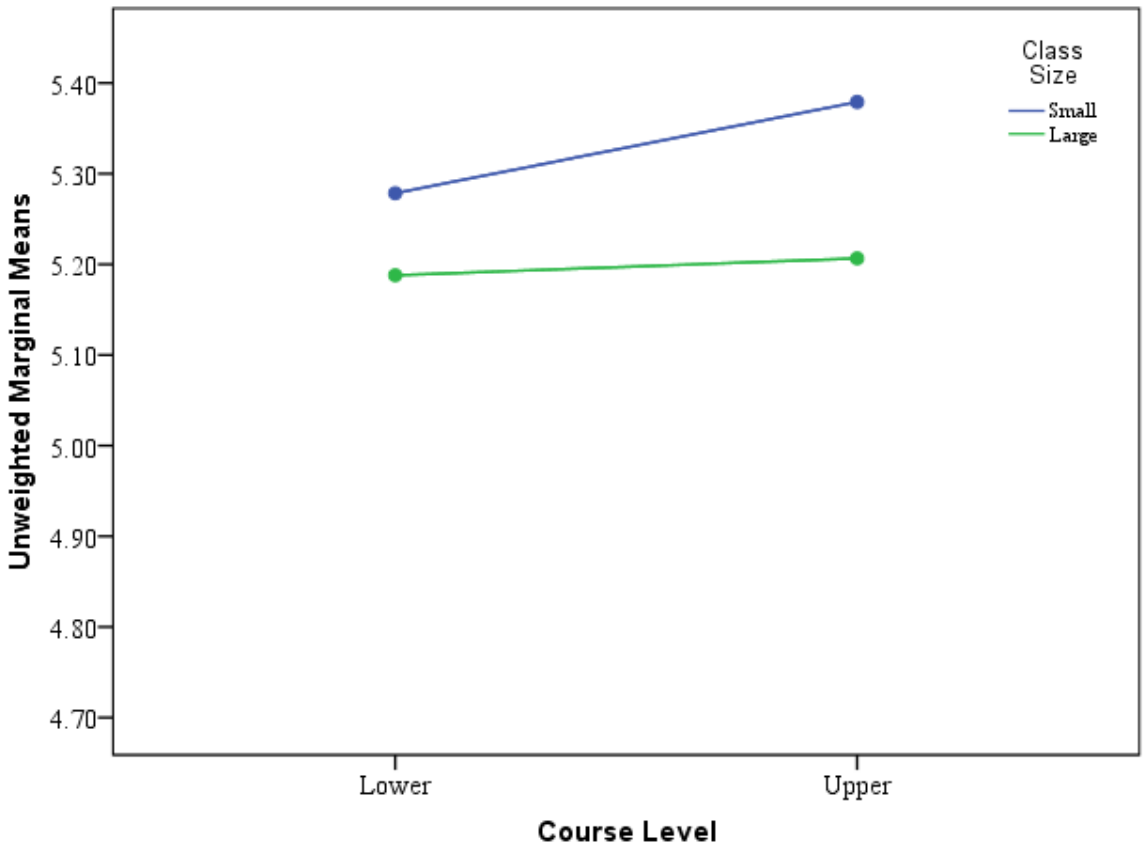

Figure 12. Plot of Equally Weighted Marginal Means for Ratings in the Artistic Environment: Class Size by Course Level.

In the social environment, there was a non-significant main effect of class size, $F(1, 243) = .24$, $p = .625$, $\eta^2 = .001$, a non-significant main effect of course level, $F(1, 243) = 1.25$, $p = .264$, $\eta^2 = .005$, and a non-significant interaction, $F(1, 243) = .33$, $p = .567$, $\eta^2 = .001$. Table 25 shows the means and confidence intervals for ratings in this environment. Figure 13 shows the graphic depiction of the interaction.

In the realistic environment, there was a significant main effect of class size, with small classes receiving higher ratings than large classes, $F(1, 333) = 5.93$, $p = .015$, $\eta^2 = .017$, and a significant main effect of course level, with lower level classes receiving higher ratings than upper level classes, $F(1, 333) = 12.56$, $p < .001$, $\eta^2 = .035$. The interaction was not significant, $F(1, 333) = 2.08$, $p = .150$, $\eta^2 = .006$. Table 26 shows the means and confidence intervals for ratings in this environment. Figure 14 shows the graphic depiction of the interaction.

In the enterprising environment, there was a non-significant main effect of class size, $F(1, 665) = .53$, $p = .468$, $\eta^2 = .001$, and a non-significant main effect of course level, $F(1, 665) = .18$, $p = .673$, $\eta^2 < .001$. However, the interaction was significant, $F(1, 665) = 4.19$, $p = .041$, $\eta^2 = .006$. Table 27 shows the means and confidence intervals for ratings in this environment. Figure 15 shows the graphic depiction of the interaction.

In the investigative environment, there was a significant main effect of class size, with small classes receiving higher ratings than large classes, $F(1, 2457) = 11.74$, $p = .001$, $\eta^2 = .005$, and a significant main effect of course level, with upper level classes receiving higher ratings than lower level classes, $F(1, 2457) = 6.08$, $p = .014$, $\eta^2 = .002$. The interaction was not significant, $F(1, 2457) = 1.61$, $p = .204$, $\eta^2 = .001$. Table 28 shows the means and confidence intervals for ratings in this environment. Figure 16 shows the graphic depiction of the interaction.

Table 25

Equally Weighted Means and Confidence Intervals for Ratings in the Social Environment (n = 247): Class Size by Course Level

Class Size	Course Level		Marginal
	Lower	Upper	
Small			
Mean (*SE*)	5.07 (.07)	5.22 (.08)	5.14 (.05)
95% CI	[4.92, 5.20]	[5.04, 5.37]	[5.03, 5.25]
Large			
Mean (*SE*)	5.08 (.09)	5.12 (.10)	5.10 (.07)
95% CI	[4.89, 5.24]	[4.90, 5.32]	[4.96, 5.23]
Marginal			
Mean (*SE*)	5.07 (.06)	5.17 (.06)	5.12 (.04)
95% CI	[4.96, 5.18]	[5.04, 5.30]	[5.04, 5.20]

Note. Standard errors and confidence intervals are based on 10,000 bootstrap samples for the mean. CI = bias corrected and accelerated confidence interval for the mean.

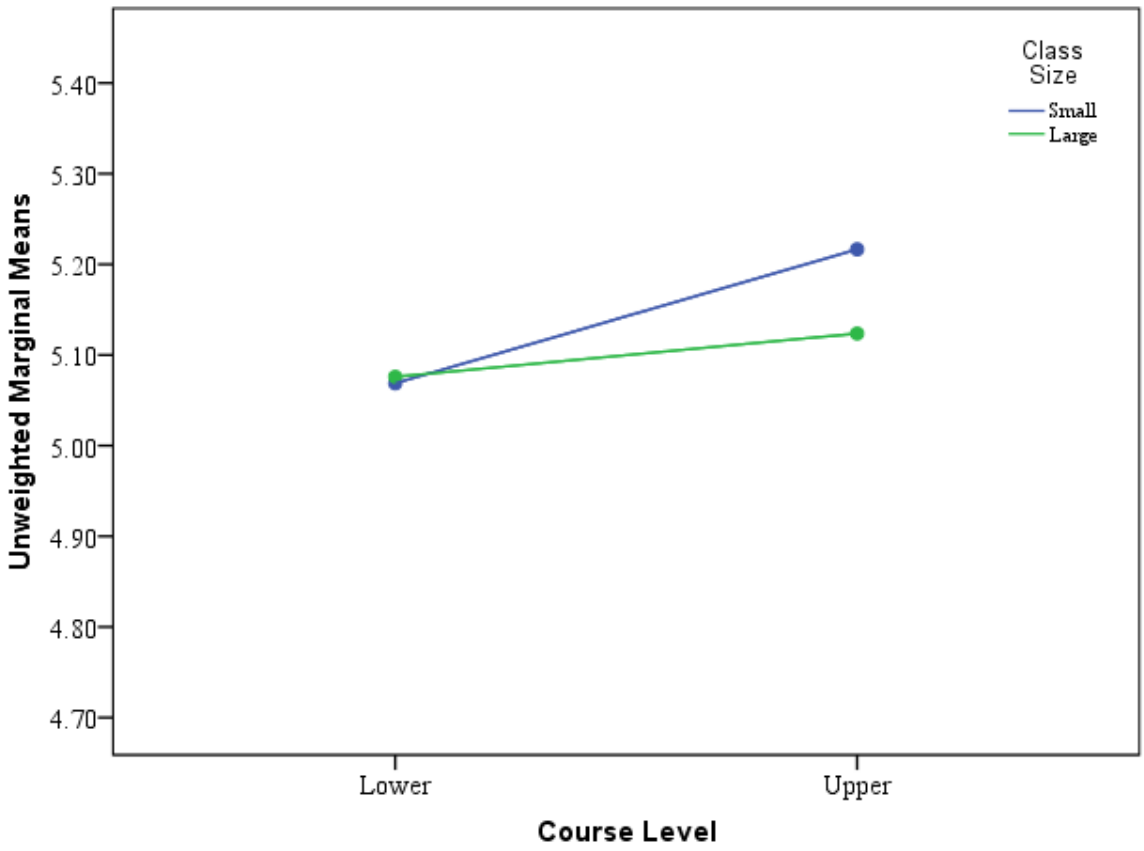

Figure 13. Plot of Equally Weighted Marginal Means for Ratings in the Social Environment: Class Size by Course Level.

Table 26

Equally Weighted Means and Confidence Intervals for Ratings in the Realistic Environment (n = 337): Class Size by Course Level

Class Size	Course Level		Marginal
	Lower	Upper	
Small			
Mean (*SE*)	5.29 (.06)	5.12 (.06)	5.21 (.04)
95% CI	[5.17, 5.41]	[5.01, 5.23]	[5.12, 5.29]
Large			
Mean (*SE*)	5.21 (.11)	4.81 (.08)	5.01 (.07)
95% CI	[4.95, 5.42]	[4.64, 4.98]	[4.86, 5.15]
Marginal			
Mean (*SE*)	5.25 (.06)	4.97 (.05)	5.11 (.04)
95% CI	[5.11, 5.37]	[4.87, 5.07]	[5.02, 5.19]

Note. Standard errors and confidence intervals are based on 10,000 bootstrap samples for the mean. CI = bias corrected and accelerated confidence interval for the mean.

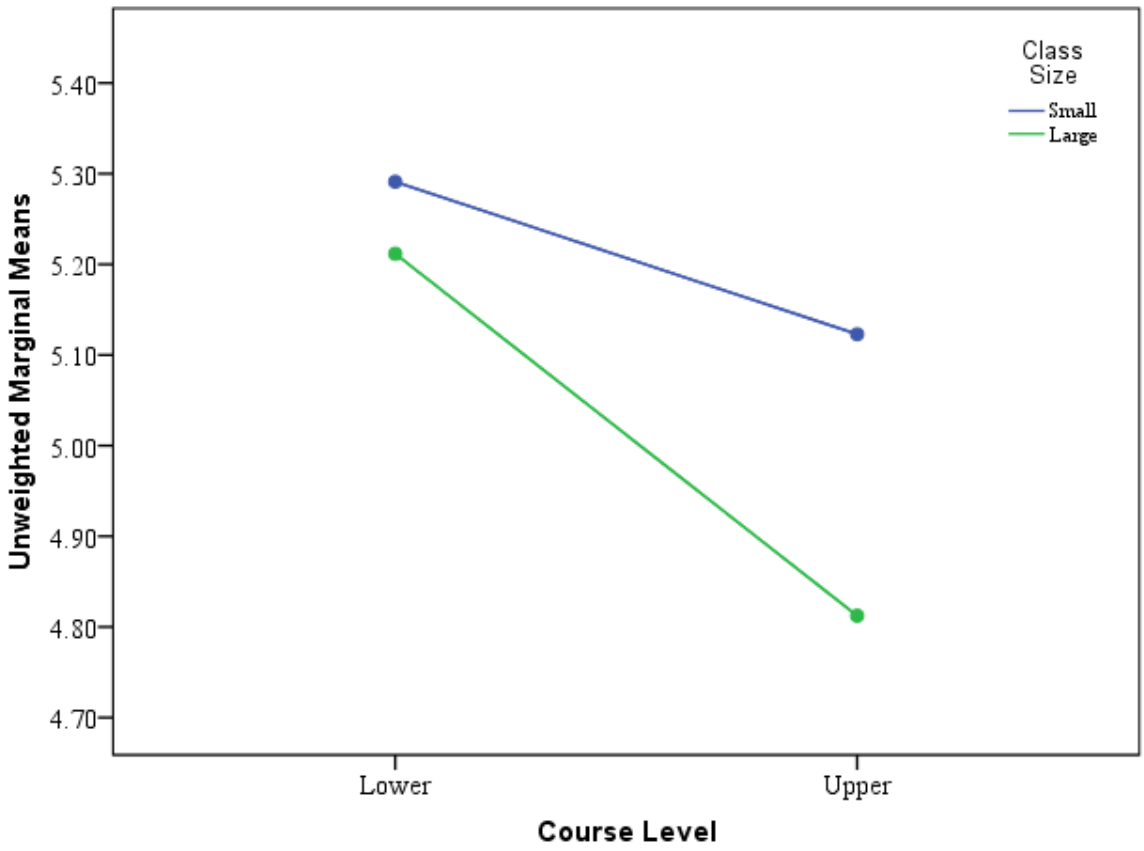

Figure 14. Plot of Equally Weighted Marginal Means for Ratings in the Realistic Environment: Class Size by Course Level.

Table 27

Equally Weighted Means and Confidence Intervals for Ratings in the Enterprising Environment (n = 669): Class Size by Course Level

	Course Level		
Class Size	Lower	Upper	Marginal
Small			
Mean (*SE*)	5.06 (.05)	5.14 (.04)	5.10 (.03)
95% CI	[4.96, 5.15]	[5.06, 5.21]	[5.04, 5.16]
Large			
Mean (*SE*)	5.12 (.06)	5.00 (.05)	5.06 (.04)
95% CI	[5.01, 5.23]	[4.89, 5.10]	[4.98, 5.14]
Marginal			
Mean (*SE*)	5.09 (.04)	5.07 (.03)	5.08 (.03)
95% CI	[5.01, 5.16]	[5.00, 5.13]	[5.03, 5.13]

Note. Standard errors and confidence intervals are based on 10,000 bootstrap samples for the mean. CI = bias corrected and accelerated confidence interval for the mean.

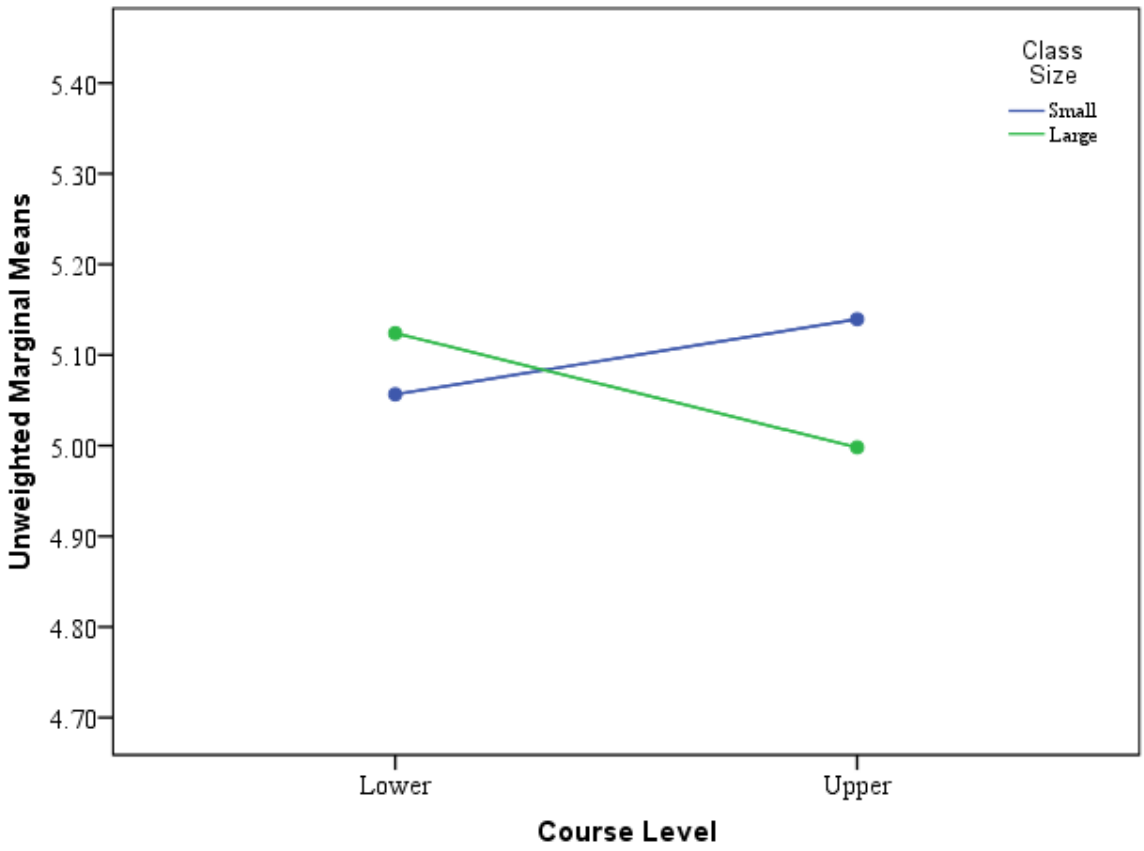

Figure 15. Plot of Equally Weighted Marginal Means for Ratings in the Enterprising Environment: Class Size by Course Level.

Table 28

Equally Weighted Means and Confidence Intervals for Ratings in the Investigative Environment (n = 2,461): Class Size by Course Level

	Course Level		
Class Size	Lower	Upper	Marginal
Small			
Mean (*SE*)	4.89 (.02)	5.01 (.03)	4.95 (.02)
95% CI	[4.85, 4.94]	[4.95, 5.07]	[4.91, 4.99]
Large			
Mean (*SE*)	4.83 (.03)	4.86 (.04)	4.85 (.02)
95% CI	[4.77, 4.88]	[4.78, 4.94]	[4.80, 4.89]
Marginal			
Mean (*SE*)	4.86 (.02)	4.94 (.03)	4.90 (.02)
95% CI	[4.82, 4.90]	[4.89, 4.98]	[4.87, 4.93]

Note. Standard errors and confidence intervals are based on 10,000 bootstrap samples for the mean. CI = bias corrected and accelerated confidence interval for the mean.

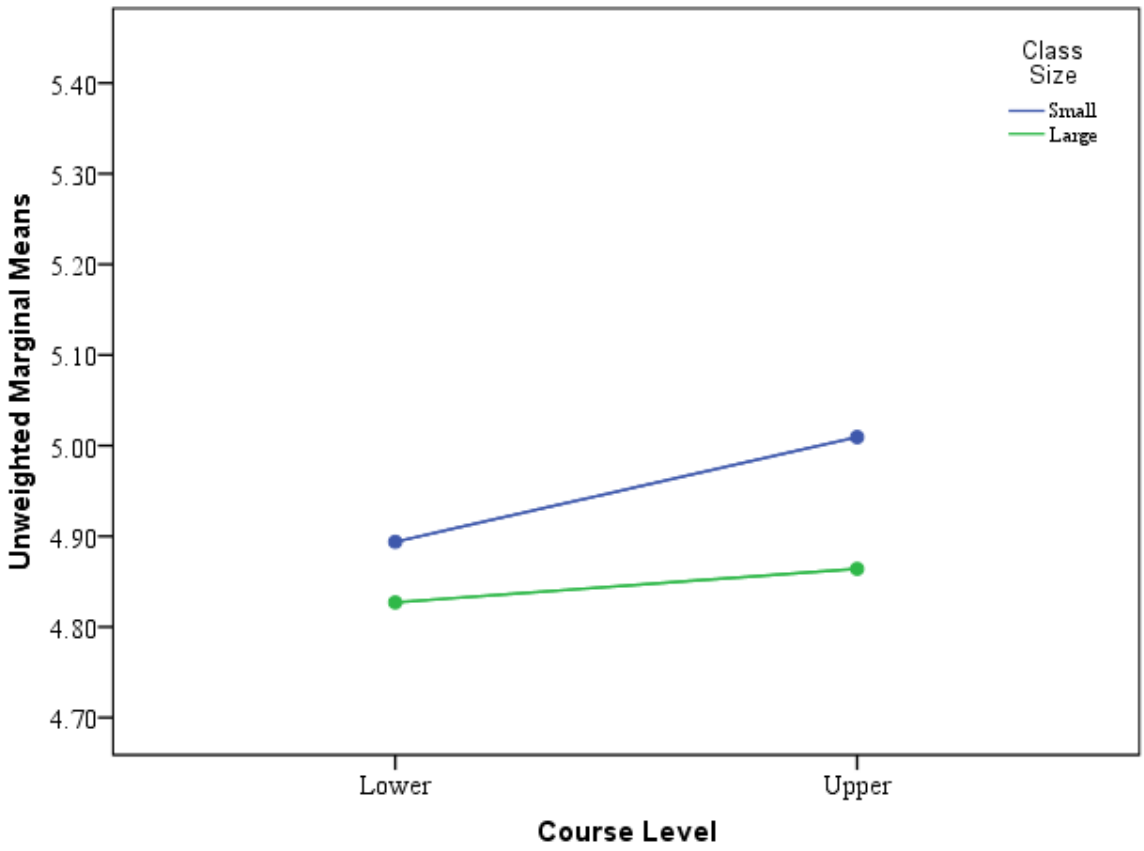

Figure 16. Plot of Equally Weighted Marginal Means for Ratings in the Investigative Environment: Class Size by Course Level.

Comparing Academic Fields

The seventh research question asks which academic field definition results in more internally consistent groups: those based on the University's academic college system or those based on Holland's theory of academic environments. To answer this question, the eta squared values from the one-way ANOVA for academic college and the one-way ANOVA for Holland environment were compared.

Eta squared (η^2), an effect size measure, represents the proportion of variance explained by each classification method. Results generated for research questions one and two show that eta squared for academic college (η^2= .067) is larger than eta squared for Holland environment (η^2= .055), suggesting that academic colleges account for more of the total variance than Holland environments[11]. However, it is possible that the small difference between these two eta squared values was due to random variation rather than the effect of classification method.

To determine the statistical significance of the eta squared difference, a bootstrap procedure was used. A bootstrap resampling was performed within the cells of a two-way table that crossed the two classification methods. Eta squared was then calculated for each classification method by conducting one-way ANOVA on the resampled data. These values were used to estimate the properties of the sampling distribution (mean = .013, standard error = .005). By sorting the values and finding the 2.5th and 97.5th percentiles, the bootstrap 95% confidence interval was estimated [.004, .022]. The confidence interval does not include zero, which indicates that the difference between classification method effects is unlikely to have occurred by chance. In other words, the difference in variance accounted for by academic college and Holland environment is

[11] The effect size for academic college is medium and the effect size for Holland environment is small according to Cohen's (1988, pp. 285 - 287) rules of thumb: η^2 = .010 is small, η^2 = .059 is medium, η^2 = .138 is large.

statistically significant. Classification by academic college resulted in slightly more internally consistent groups than classification by Holland environment. Figure 17 shows the bootstrap sampling distributions for eta squared when ratings are classified by Holland environment and by academic college. Figure 18 shows the bootstrap sampling distribution for the eta squared difference. Additional details about the bootstrap procedure are in Appendix D.

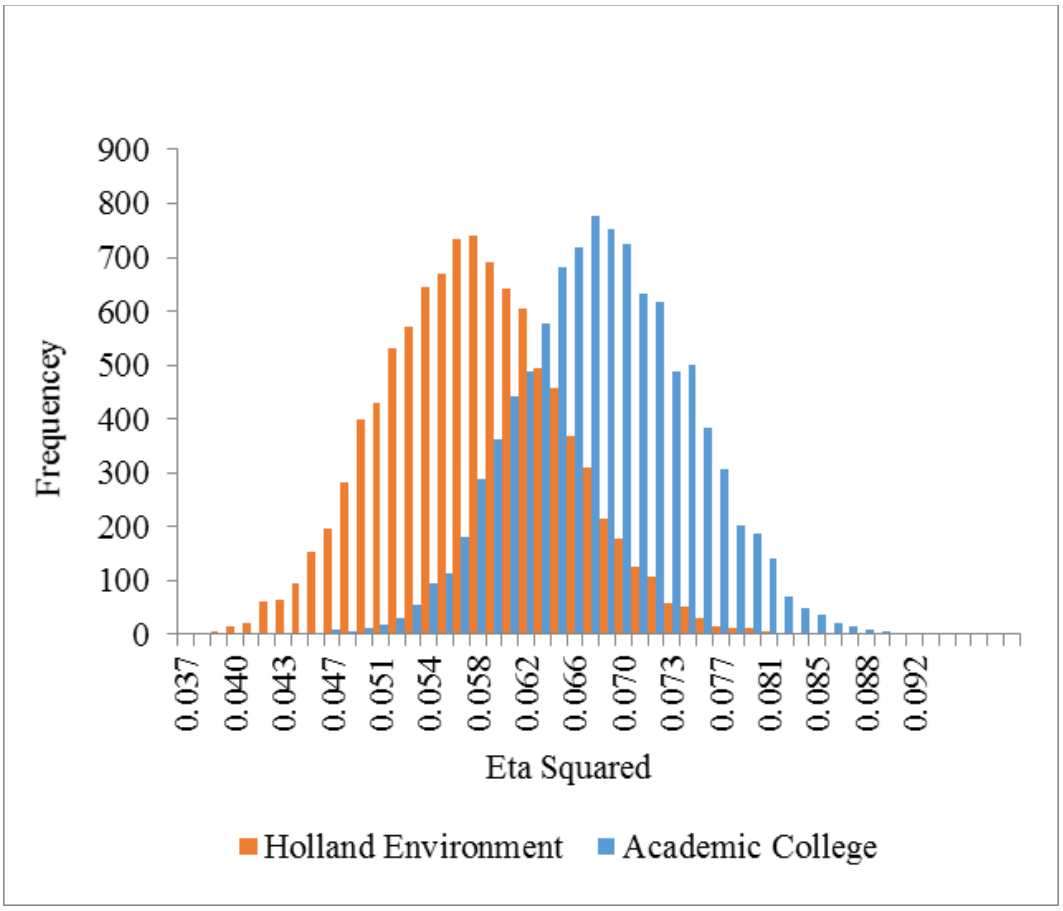

Figure 17. Bootstrap Sampling Distributions for Eta Squared When Ratings are Classified by Holland Environment and by Academic College.

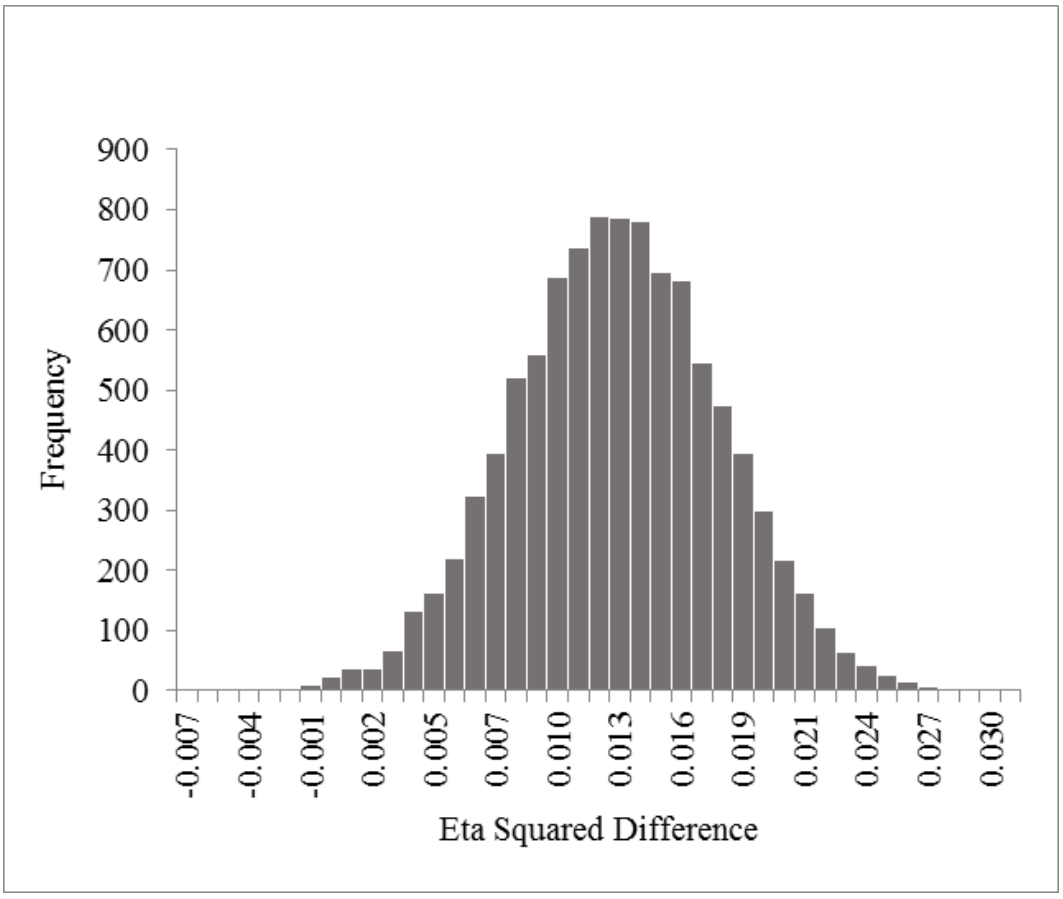

Figure 18. Bootstrap Sampling Distribution for the Difference between Eta Squared for Holland Environment and Academic College (η^2 environment - η^2 college)

Chapter Five

Discussion

This study was designed to examine relationships between student ratings of instruction, and the academic field, level, and size of a class. The data were mean class ratings collected from more than 4,000 class sections at a large research university. Analysis of variance was used to answer seven research questions that queried the role of each independent variable, interactions among the independent variables, and the comparative effects of two approaches to categorizing classes into academic fields. This chapter discusses the meaning of the results and their relationship to prior findings. Implications for future practice, policy, and research are also noted.

Academic Colleges and Holland Environments

The first and second research questions ask whether differences in ratings are associated with academic colleges or Holland environments, respectively. Results indicated that both classification schemes accounted for significant differences between academic fields. The effect size for college was $\eta^2 = .067$ (i.e., 6.7% of the rating variance was associated with academic college). The effect size for environment was $\eta^2 = .055$ (i.e., 5.5% of the rating variance was explained by Holland environment). The effect size for college is medium and the effect size for environment is small according to Cohen's (1988) rules of thumb for η^2, which indicates these findings have practical as well as statistical significance.

A context for understanding these findings is provided by two bodies of research. Studies of disciplinary differences in student ratings provide context for the results of classification by academic college. Although no past research has examined student ratings within the framework of Holland's theory, there are studies of teaching practices within Holland environments, so this

research provides a perspective on the results of classification by Holland environment.

Results from the classification by academic college show that ratings were highest in the colleges of Liberal Arts and Human Sciences, and Agriculture and Life Sciences; lowest in Science, and Engineering; and intermediate in Architecture and Urban Studies, and Business. The order of mean ratings is consistent with past research. For example, Feldman (1978) conducted a comprehensive review of the student ratings literature and ranked the ratings in different academic fields. He concluded that English, the humanities, the arts, and foreign languages tend to be in the high and medium rankings. Social sciences, such as political science, sociology, psychology, and economics, tend to be in the medium or low rankings. Science fields (except certain biological sciences) mathematics, and engineering tend to be in the lowest rankings. Years later, Cashin (1990b), and Franklin and Theall (1995) reached very similar conclusions.

There are a number of possible explanations for the difference in ratings between academic colleges. One is that more quantitatively oriented courses may receive lower ratings. Past research shows that greater student achievement is related to higher ratings (Cohen, 1981) and that students will often attribute their poor academic performance to factors external to themselves (Theall, Franklin, & Ludlow, 1990). Cashin (1990b) argued that students' quantitative skills are less developed than their verbal skills; thus, they may experience dissatisfaction with poor performance in quantitative classes and this may be reflected in ratings. Quantitatively oriented courses may also be more difficult to teach well, as a result of lower student preparedness and/or the nature of the subject matter. This explanation would account for the low ratings in Science, Engineering, and Business but it does not account for the high ratings in Agriculture and Life Sciences - where subjects such as Agricultural Economics, Biochemistry,

Crop and Soil Science, and Food Science all have significant quantitative components. It may be that natural science courses are an exception to what is an otherwise strong pattern, and that they receive high ratings for other reasons, despite their quantitative orientation.

Another possible explanation for the differences in ratings across academic colleges is related to research showing that ratings are associated with students' sense of involvement in the course (Remedios & Lieberman, 2008). Gillmore (1994) found that the more students' value the time spent preparing for class, the higher the student ratings of instruction. Centra (2003) found that time spent on coursework could be divided into "good" hours (i.e., deemed valuable by students) and "bad" hours. Franklin and Theall (1995) referred to the ratio of these variables as "time-valued", and they explored the relationship between this ratio and students' ratings of instruction. They found that the amount of ratings variation explained by the time-valued ratio varied widely across academic fields. It accounted for the largest proportion of variation in mathematics, engineering, and the sciences. It accounted for the smallest proportion of variation in the humanities. The arts, agriculture, business, and the social sciences fell in-between in terms of the proportion of explained variance. This ordering of academic field aligns with the rank order of mean college ratings in the present study. The effect of the time-valued ratio tends to be larger in the colleges with the lowest ratings (Science and Engineering) and smaller in the college with the highest ratings (Liberal Arts and Human Sciences, which includes the humanities). Thus, it is reasonable to think that the order of mean college ratings may reflect students' perceptions of the value of assignments and coursework completed outside of class time, as well as in class activities.

A final possible explanation is that courses in the colleges with lower ratings tend to be more poorly taught. Teaching methods are influenced by the culture of the academic discipline,

department, and college; some cultures are likely to support high quality teaching more so than others. Franklin and Theall (1992) compared the instructional goals and teaching methods of instructors in humanities, business, and science and engineering. They found that arts and humanities instructors - to a greater extent than instructors in other academic fields - tended to emphasize “thought” goals more so than “fact” goals and to use discussion and independent projects rather than lecturing alone. Since these teaching practices independently correlate with student ratings they may account for the higher ratings in the Liberal Arts and Human Sciences. Another reason that courses in certain colleges may be more poorly taught is related to finances. In Science and Engineering, the University must pay high salaries to compete with business and industry. It is easier to find, and less expensive to hire, an outstanding teacher in English, music, philosophy, history, or agriculture, than in engineering, computer science, or business.

Results from the classification by Holland environments show that ratings were highest in the artistic environment (example subjects include English, foreign languages the arts, architecture, and philosophy); lowest in the investigative environment (example subjects include mathematics biochemistry, accounting information systems, and engineering); and intermediate in social, realistic, and enterprising environments. These results are consistent with prior research showing that arts and humanities courses tend to receive higher ratings than math, science, and engineering courses (Cashin, 1990b; Feldman, 1978; Franklin & Theall, 1995). Course ratings in the social environment were slightly higher than in the realistic and enterprising environments (an outcome that past research would suggest); however, the difference between the means was not statistically significant. The lack of significant differences between the means may be due to the particular mix of subjects within these environments. Subjects in the realistic environment include animal sciences and horticulture (both tend to receive high rating), as well as mechanical

engineering and building construction (among those that tend to receive the lowest ratings). Subjects in the enterprising environment include communication studies (tends to receive high ratings), as well as finance and marketing (tend to receive low ratings).

Possible explanation for the differences that were found between environments comes from a basic premise of Holland's theory that academic environments reinforce and reward students and faculty for the display of attitudes, interests, and abilities that are consistent with the values of their particular academic field. Particularly relevant to understanding the present findings are the relationships found between Holland environments and the ways in which faculty members structure their courses and interact with students, and the educational goals they emphasize in their teaching (Smart, 1982; Smart & Thompson, 2001; Smart & Umbach, 2007; Thompson & Smart, 1999). In general, faculty in artistic and social environments are more likely than faculty in other environments to use educational practices that engage students in active and collaborative learning and student-faculty interaction (Umbach, 2006). Also, in their courses, faculty from artistic and social environments place a greater emphasis on understanding people as compared to their peers in investigative and enterprising environments, who place greater emphasis on analyzing data (Smart & Umbach, 2007). It is likely that the activities and goals emphasized in the artistic and social environments may lead to greater levels of student satisfaction and engagement, which in turn may lead to higher student ratings.

Following the finding of significant differences between college means and between environment means, analyses turned to differences in the variance of ratings within groups. The variance within a group matters when raw class means are being compared (e.g., for teaching awards or merit pay decisions). If there is little variance among the ratings within a group, than even a small difference between scores may be interpreted as meaningful; conversely, the same

difference would be less meaningful if the ratings spread widely across the response scale.

Within the classification by college design, variance differed significantly between groups, with the highest variance occurring in the College of Science and the lowest variance occurring the in the College of Liberal Arts and Human Sciences. This was an unexpected finding because the College of Liberal Arts and Human Sciences houses a greater number and diversity of subjects than the College of Science and would, therefore, be expected to have greater variance in class ratings. A plot of variance against the mean for each college showed that variance tends to become smaller as the group mean increases. A plausible explanation for this pattern relates to the scale on which the ratings were collected. The dependent variable was measured on a 6-point Likert scale and the highest college means (5.24 in the College of Liberal Arts and Human Sciences and 5.23 in the College of Agriculture and Life Sciences) are close to the upper limit of this scale. Further, although rating distributions in all colleges are negatively skewed, the skew decreases from -1.58 and -1.40 in the colleges of Agriculture and Life Sciences, and Liberal Arts and Human Sciences, respectively, to -1.10 and -.74 in the colleges of Business and Engineering, respectively. In other words, greater skew accompanies higher means and lesser skew accompanies lower means. These patterns suggest that a ceiling effect is limiting the variability of ratings, with the greatest impact of this ceiling occurring in the colleges with the highest means.

Similar results were found within the classification by environment design. The ratings variance differed significantly across environments, with the greatest variance occurring in the investigative environment and the smallest variance occurring in the artistic environment. As was suggested previously, an explanation for this finding may be the scale on which the ratings were collected. As was the case with the classification by college design, the highest group

mean (5.30 in the artistic environment) is close to the upper limit of the 6-point response scale. Additionally, the skew of the rating distributions decreases from -1.45 and -1.55 in the artistic and social environments, respectively, to -1.17 and -1.08 in the enterprising and investigative environments, respectively. Thus, greater skew accompanies the higher means while lesser skew accompanies the lower means. These patterns suggest that a ceiling effect may be limiting variability, especially as mean ratings approach the upper limit of the response scale.

The variability findings provide an additional reason for interpreting student ratings within appropriate norm groups because, in addition to differing in terms of the mean across academic fields, ratings may differ in terms of their variability across academic fields. While this finding is linked to the particular response scale used in this study, ratings results are likely to be skewed toward the high end of the response scale in most instruments. This suggests that absolute differences between class ratings cannot be interpreted in the same way across academic fields.

Course Level

The third research question asked whether differences in ratings are associated with course level. Results showed significantly higher ratings in upper level classes than in lower level classes. This was the expected result based on prior research finding the same (Aleamoni, 1981; Braskamp & Ory, 1994; Feldman, 1978). However, the small difference between the means (.07) is inconsequential for most practical purposes. The effect size was $\eta^2 = .003$ (i.e., .3% of the rating variance was explained by course level). This effect size is trivial according to Cohen's (1988) rules of thumb. In addition, variance did not differ significantly between the two course levels. These findings suggest that upper and lower course divisions, on their own, may play only a minor role in the differences between student ratings.

Class Size

The fourth research question asked if differences in ratings are associated with class size. Results showed significantly higher ratings and less variance in small classes as compared to large classes. The effect size was $\eta^2 = .014$ (i.e., 1.4% of the rating variance was explained by class size). This is a small (but still meaningful) effect according to Cohen's (1988) rules of thumb. The difference between the means (.17) would have practical significance in many contexts (the standard deviation of all ratings is .69).

This finding is consistent with prior research showing that class size has a small but significant effect on ratings (Feldman, 1984; Hoyt & Lee, 2002a). While smaller classes tend to receive higher ratings, students in those classes also score higher on measures of self-reported learning for course outcomes (Centra, 2009). Thus, researchers have argued that class size should not be considered a biasing factor because evidence suggests that students in smaller classes actually do learn more. Still, these findings indicate that it is appropriate to take class size into consideration by using appropriate comparative data to interpret student ratings (Benton & Cashin, 2012; Centra, 2009).

Academic College x Course Level x Class Size

The fifth research question asked whether there are interaction effects between academic college and course level or class size. Results showed a significant main effect of college and class size, and a non-significant main effect of course level. However, the three-way interaction was significant, indicating that the main effects should be interpreted with caution because interaction among the three variables is associated with differences between the ratings.

Two themes emerged from follow up analyses of the interaction of class size and course level within each college. First, the interaction between class size and course level differs across

colleges, suggesting that the relationship between these two variables was not the same in each college. In the majority of colleges there was no interaction, so the effect of class size could be interpreted independently from the effect of course level. In two of the six colleges, however, there were interaction effects between class size and course level. In the College of Science, the interaction was an exaggeration of expected effects; in the College of Business, the interaction was unexpected.

In the College of Science, in lower level courses, there was little difference between ratings for small versus large classes; however, in upper level Science courses there was a considerable difference, with small classes receiving higher ratings than large classes. Put differently, although small classes receive relatively higher ratings at both course levels, the difference in ratings between small and large classes is greater in upper level courses. This result is consistent with the expectation that students give higher ratings to small classes and is, therefore, not surprising.

In the College of Business, in lower level courses, students rate small classes higher than large classes. This is the expected result. Contrasting this, in upper level courses, ratings for small classes are slightly lower than ratings for large classes. This is surprising. While ratings for large classes barely differ between the lower and upper levels, the ratings for small classes decrease considerably, from a mean of 5.32 in lower level courses to a mean of 4.98 in upper level courses. Although this interaction is difficult to explain, it does suggest that smaller classes are not always better from a student's perspective. Further research is needed to determine what other variables may be associated with the unique interaction effect in this particular college.

A second theme in these results is the differing proportion of variance accounted for by class size and course level across colleges. This is consistent with prior research finding that the

correlations between class size, course level, and student ratings are not the same in all academic fields (Franklin & Theall, 1995). In this study, the class size and course level had differing effects on student ratings across colleges. One or both of the variables of interest had a significant effect on ratings in every college except Engineering, where neither variable had a significant effect on ratings. Looking only at results that show significant effects, the amount of variance accounted for by either of these variables ranges from about .3% (in the College of Liberal Arts and Human Sciences) to about 13% (in the College of Science). This implies that the impact of these variables on student ratings is best considered within academic fields. For example, when interpreting ratings, class size may be an important factor in the College of Science but not in the College of Engineering.

Holland Environment x Class Size x Course Level

The fifth research question asked whether there are interaction effects between Holland environments and course level or class size. Results showed a significant main effect of environment, although there was also a significant interaction between environment and course level, indicating that the effect of environment is moderated by course level. There was a significant main effect of class size, although there was also a significant interaction between class size and course level, indicating that the effect of class size is moderated by course level. The main effect of course level was not significant, nor was the interaction between environment and class size. The three-way interaction was not significant, indicating that Holland environment, class size and course level made unique contributions to the student ratings. Follow-up tests were conducted to facilitate understanding of the interaction effects of class size and course level within each environment.

The themes that emerged from the follow up analyses were very similar to those that

emerged in the College x Class Size x Course Level Design. First, the interaction between class size and course level differs across environments, suggesting that the relationship between these variables is not the same in each environment. In four of the five environments there was no interaction, so the effect of class size could be interpreted independently from the effect of course level. In the enterprising environment there was an interaction effect between class size and course level.

In the enterprising environment, in lower level courses, students rate large classes higher than small classes. This contradicts the expectation of many faculty members that small classes receive higher ratings. It is surprising that the large class advantage occurs in lower level courses, which typically enroll a greater percentage of first- and second- year students. In upper level courses, the results were as expected, with small classes receiving higher ratings than large classes. Looking across course levels within the enterprising environment, from lower to upper, it is apparent that small class ratings increase about the same amount that large class ratings decrease. More research would likely reveal other influential variables that are associated with the unique interaction in this environment. It should be noted that courses in the enterprising environment account for 58% of the courses in the College of Business (i.e., the only college where an interaction effect was found between class size and course level) – so many of the same classes are contributing to the interaction effect found in both of these cases.

Comparing Academic Fields

The seventh research question asked which academic field definition results in more internally consistent groups: those based on the University's academic college system or those based on Holland's theory of academic environments. Eta squared for academic college (η^2= .067) was found to be larger than eta squared for Holland environment (η^2= .055), indicating that

the academic colleges accounted for more ratings variance than the Holland environments (η^2 difference = .012, a difference of 1.2% of total ratings variance). A bootstrapping procedure was used to test this difference and results indicated that the difference between the classification method effects was unlikely to have occurred by chance.

In this study of student ratings, Holland's theory of academic environments did not result in more internally consistent academic fields when compared to the academic fields defined by the University's own academic college system. Perhaps this result is not surprising if, as prior research suggests (Umbach, 2007), what is taught and how it is taught is shaped by the interaction of institutional and disciplinary culture. In American research universities, academic colleges organize disciplinary departments in ways that are unique to the institution. As a result, disciplinary cultures and institutional cultures overlap within each academic college. In contrast, Holland's theory addresses disciplinary, but not institutional, culture. Given this, a possible explanation for why academic colleges accounted for more ratings variance than Holland environments is that the academic colleges combine institutional and disciplinary effects.

Although the academic colleges created more internally consistent groups, Holland environments were also associated with significant differences in ratings and the proportion of variance accounted for by academic college and by Holland environment was similar. The subject groupings created by the two classification methods differed enough to suggest that each contributed something unique to the definition of academic fields. Further, the Holland environments approach to classification resulted in a clean model (without three way interaction). The lack of interaction means that Holland environment, class size and course level made unique contributions to the student ratings. More research is needed to explain the relationships between Holland environments and student ratings, but the findings in this study

suggest that Holland environments may be useful in defining academic fields for the interpretation of student ratings. This may be especially true when subjects that tend to receive high ratings are located in the same academic college as subjects that tend to receive low ratings. The particulars regarding how to best use Holland's theory for this purpose could be researched within individual institutions.

Limitations

As with all research, this study had limitations that should be considered along with the findings. A few of the major limitations are mentioned here.

The most obvious limitation is generalizability. The data for this study were collected at one university in a single academic year. It is likely that some of the findings would have been different if this study were conducted at a different institution of higher education, or during a different time frame at the same institution.

The study examined the influence of particular course variables and did not include student or instructor variables. Any of the variables not included in the study could potentially influence ratings or interact with the variables in this study to change the outcome.

Many alternative methods for categorizing classes into academic fields were not included in this study. Rather, the study focused on comparing two different methods for defining academic fields. It is also possible that there are more influential course characteristics (other than academic field), such as the amount of quantitative work, or the distinctions between hard and soft fields of study.

The Educational Opportunities Finder (Rosen, Holmberg, & Holland, 1997) applies Holland's theory to the classification of more than 900 college fields and was used to guide the process of categorizing classes into Holland environments. Although this process was

straightforward in most cases, there were several subjects where the classification indicated by the Educational Opportunities Finder was not as expected. For example, the following subjects did not intuitively fit with the academic environment indicated: history, and political science (enterprising); sociology (investigative); materials science and engineering, and mechanical engineering (realistic). Further, subjects such as architecture, accounting and information systems, and engineering education include interdisciplinary content that is not easily classified in any one of the Holland environments. An update to Holland's theory and/or the Educational Opportunities Finder classifications may be needed, especially as interdisciplinary collaborations and majors are becoming more common. Updates could potentially make Holland's theory more useful in student ratings research and in student ratings interpretation.

Finally, when categorizing classes into Holland environments, the course subject code was used to identify the subject matter for the class. The subject code is a general identifier for the subject of the course and does not account for cross-disciplinary courses or courses taught outside their disciplinary home department. More accurate identification of the subject matter for each classes may have changed the results.

Implications

The results of this study are consistent with prior research findings and they also offer new knowledge about the role of course variables in student ratings of instruction. The findings have possible implications for individuals who use student ratings in personnel decisions, to improve their teaching, or as the subject of their research.

A clear conclusion to emerge from this study is that the variation in ratings across academic fields makes it inadvisable to compare instructors or courses outside of appropriate norm groups. Ideally, norm groups should be homogeneous enough that they exclude much of

the variation caused by differences in course variables such as academic field, class size, and course level. At the same time, in a single academic year, the norm groups should contain enough classes to protect the confidentiality of individual instructors and provide a reliable basis for the comparison of ratings. This means that each norm group should contain enough classes that one or two outliers will not dramatically affect the overall mean.

Appropriate norm groups could be created by defining "like" groups of classes – that is, classes in similar academic fields (however defined), at the same course level, and with approximately the same enrollment. Information about students' motivation for taking a class could also be used to group classes based on the students' level of prior interest in the subject matter. Analysis of rating means and variability - within and between the norm groups – could inform subsequent revisions and improvements to the groups. Ideally, this process would be iterative, with changes made over time to reflect changing circumstances.

This study indicates that norm groups could be created by using multiple approaches to the definition of academic fields. Organizational structures, such as an institutions' own academic college system, may be the best place to start. Holland environments could provide a useful compliment to institution-specific definitions of academic fields, especially when an institution's academic colleges are too small, too large, or too diverse (in terms of the academic disciplines they contain) to provide appropriate reference groups.

In the institution where this study was conducted, the College of Liberal Arts and Human Sciences provides an example situation where an academic college may be too diverse to serve as an appropriate reference group. The College of Liberal Arts and Human Sciences, and Holland's artistic environment were the two groups in this study that had the highest mean ratings when compared to other groups in their respective classification schemes. Not

surprisingly, many of the same classes are in these two groups: 734, or 52% of classes in the College of Liberal Arts and Human Sciences are categorized in the artistic environment (see Table 5 in Chapter 3). These classes account for the most highly rated academic fields in both approaches to defining academic field. In contrast, 89, or 6% of classes in the College of Liberal Arts and Human Sciences are categorized in the investigative environment, which had the lowest mean rating when compared to other groups in the Holland scheme. Although ratings of the 89 investigative classes are likely to be below average when compared to others in the College of Liberal Arts and Human Sciences, these same classes are likely to be above average when compared to others in Holland's investigative environment. In this scenario, the academic college that houses these courses probably provides an inappropriate reference group. It may be more appropriate to consider these 89 courses as norm group unto themselves, or they could be compared to other courses that are in the investigative environment (even if those courses are in different academic colleges).

Instructors may benefit from improved information about appropriate norm groups and inappropriate comparisons between student ratings in different classes. One way that faculty members make sense of the student ratings they receive is by comparing them to ratings received by colleagues in their department, college, and institution. However, some of these comparisons may be misleading. Knowing how class size and course level tend to influence ratings within their particular academic field may help instructors to interpret student ratings in more appropriate contexts and guide teaching improvements. The best way to share this information with instructors may be by changing the way results are reported. If appropriate comparative data are provided to instructors along with their ratings results, the likelihood of inappropriate comparisons between different classes would be reduced. Findings from this study also indicate

that measures of the variability of ratings (within the norm group) are important for making sense of differences between the ratings received in different class sections. This information could be effectively communicated with visuals such as box plots or histograms, which are easily understood by individuals without knowledge of statistics.

Those who study student ratings may be interested in the results related to the potential usefulness of Holland's theory for the classification of academic fields. More research is needed to determine whether Holland's theory might provide a reasonable alternative to atheoretical or institution-specific definitions for academic fields in student ratings research. However, the results of this study provide a first step toward finding a useful conceptual framework for the definition of academic fields. If researchers are able to organize academic fields according to a common theoretical framework, it will facilitate cross-institution comparisons by defining like-fields outside of the academic structure of any one institution.

Researchers could build on the findings from this study by testing Holland's framework for explaining academic field differences in student ratings at different institutions and institutional types. Studies that compare academic field definitions explicitly for the purpose of interpreting student ratings would be most useful. Many different approaches to the definition of academic fields exist, but the most promising are probably those based on the student experience, student learning, or faculty attitudes and behaviors related to teaching.

Many questions and hypotheses about the differences in ratings across academic fields have yet to be explored. For example, are ratings lower in more quantitatively oriented courses? To begin answering this question, a researcher could categorize classes based on the subject matter taught in each particular class. Even within academic programs, classes vary in the extent of quantitative focus. Classes with a strong quantitative focus (especially if taught outside of

quantitatively oriented departments) may form a norm group that crosses academic disciplines.

Future research could also build on this study by examining student characteristics alongside course characteristics. Past research indicates that student motivation and perception of workload are two variables that systematically influence ratings. Survey items could be designed to measure these variables - for example, items could ask students about their motivation for taking the course, and their perceptions of the usefulness and difficulty of class assignments. Much could be learned from a study designed to examine various combinations of student and course variables. Past research on student ratings of instruction has typically examined student variables such as age, gender, GPA, motivation, and academic major, separately from course variables. Although many studies have reported on the effects, or lack of effects, associated with individual variables, more useful information could come from research on the interaction of these variables.

Finally, additional student ratings research is needed to expand and update our understanding of the role of faculty characteristics and account for the new faculty diversity. There is little recent research on this subject. A useful and interesting study could examine contemporary relationships between student ratings and various instructor characteristics, such as age, gender, race, country of origin, and sexual orientation. Findings on the interaction effects between student characteristics and instructor characteristics would be especially interesting.

Conclusions

Generally, findings from this study reinforce the recommendations of Cashin (1999), and Benton and Cashin (2012), regarding the importance of appropriate reference groups for the interpretation of student ratings. The findings complement prior research showing that student ratings of instruction differ between academic fields. They replicate the general pattern found in

prior research that courses in arts and humanities are rated higher than courses in the natural and social sciences, which in turn, are rated higher than courses in the sciences, mathematics, and engineering. Also, findings from this study are consistent with prior research showing that rating differences are associated with class size and course level.

Beyond their consistency with prior research, findings of the present study are unique in several ways. First, they confirm the value of institution-specific approaches to defining academic fields for the interpretation of student ratings of instruction. Simultaneously, they indicate that Holland's theory of academic environments may be a useful conceptual framework for making sense of differences in ratings across academic fields. An institution's academic colleges may account for relatively more ratings variance than Holland environments, but Holland's theory may be potentially useful as a complementary approach to institution-specific or atheoretical academic field definitions.

Adding to past studies that reported differences in mean ratings across academic fields, this study found that academic fields might also differ in the variance of ratings. This has implications for the statistical analysis of ratings. It also supports the notion that variability measures, such as standard deviation, may be as important as norm group means in situations where ratings from different classes are being compared.

The study findings also indicate that it is important to look within, rather than just across, academic fields when researching variables that may impact ratings. Prior research showed differences in student ratings were associated with class size and course level. The findings of this study show that class size and course level may impact student ratings differently – in terms of interaction effects and magnitude of effects – depending on the academic field of the course.

The findings from this study may be of interest to faculty members and administrators

who use ratings to judge teaching quality for tenure, promotion, or merit decisions. Researchers who study student ratings may be interested in the methods as well as the results. Individuals who make polices related to personnel decisions, and individuals who manage student ratings systems, may also find information in this study's findings to support their work.

References

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

Abrami, P. C. (2001b). Improving judgments about teaching effectiveness: How to lie without statistics. In M. Theall, P. C. Abrami & L. A. Mets (Eds.), The student ratings debate: Are they valid? How can we best use them? (pp. 59-87). *New Directions for Institutional Research*, 109. San Francisco: Jossey-Bass.

Abrami, P. C., & d'Apollonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. In M. Theall & J. Franklin (Eds.), Student ratings of instruction: Issues for improving practice. (pp. 97-111). *New Directions for Teaching and Learning, 43.* San Francisco: Jossey-Bass.

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

Abrami, P. C., d'Apollonia, S., & Rosenfeld, S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective.* (pp. 385-445). Dordrecht, The Netherlands: Springer.

Abrami, P. C., Perry, R. P., & Leventhal, L. (1982). The relationship between student personality characteristics, teacher ratings, and student achievement. *Journal of Educational Psychology, 74*(1), 111-125.

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

Arreola, R. A. (2007). *Developing a comprehensive faculty evaluation system* (3rd ed.). Bolton, MA: Anker Publishing.

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

Barnes, L. L. B., & Barnes, M. W. (1993). Academic discipline and generalizability of student evaluations of instruction. *Research in Higher Education, 34*(2), 135-149.

Becher, T. (1981). Towards a definition of disciplinary cultures. *Studies in Higher Education, 6*(2), 109-122.

Becher, T. (1987). The disciplinary shaping of the profession. In B. R. Clark (Ed.), *The academic profession: National, disciplinary, and institutional settings* (pp. 271-304). Berkeley, CA: University of California Press.

Becher, T., & Trowler, P. (2001). *Academic tribes and territories: Intellectual inquiry and the cultures of disciplines* (2nd ed.). Philadelphia, PA: The Society for Research into Higher Education and Open University Press.

Benton, S. L., & Cashin, W. E. (2012). IDEA paper no. 50: Student ratings of teaching: A summary of research and literature. Manhattan, KS: IDEA Center.

Benton, S. L., Duchon, D., & Pallett, W. H. (2011). Validity of student self-reported ratings of instruction. *Assessment & Evaluation in Higher Education*, 1-12.

Biglan, A. (1973). The characteristics of subject matter in different academic areas. *Journal of Applied Psychology, 57*(3), 195-203.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and*

institutional performance. San Francisco: Jossey Bass.

Carrier, N. A., Howard, G. S., & Miller, W. G. (1974). Course evaluations: When? *Journal of Educational Psychology, 66*, 609-613.

Cashin, W. E. (1990a). IDEA paper no. 22: Student ratings of teaching: recommendations for use. Manhattan, KS: IDEA Center.

Cashin, W. E. (1990b). Students do rate different academic fields differently. In M. Theall & J. Franklin (Eds.), Student ratings of instruction: Issues for improving practice (pp. 113-121). *New Directions for Teaching and Learning,* 43 San Francisco: Jossey-Bass.

Cashin, W. E. (1995). IDEA paper no. 32: Student ratings of teaching: The research revisited. Manhattan, KS: IDEA Center.

Cashin, W. E. (1999). Student ratings of teaching: Uses and misuses. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions.* (pp. 25-44). Bolton, MA: Anker.

Cashin, W. E. (2003). Evaluating college and university teaching: Reflections of a practitioner. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 531-593). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Cashin, W. E., & Downey, R. G. (1992). Using global student ratings for summative evaluation. *Journal of Educational Psychology, 84*(4), 563-572.

Cashin, W. E., & Downey, R. G. (1995). Disciplinary differences in what is taught and in students' perceptions of what they learn and of how they are taught. In N. Hativa & M. Marincovich (Eds.), Disciplinary differences in teaching and learning: Implications for practice (pp. 81-92). *New Directions for Teaching and Learning,* 64. San Francisco, CA: Jossey-Bass.

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. 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*(5), 495-518.

Centra, J. A. (2009). Differences in responses to the Student Instructional Report: Is it bias? Princeton, NJ: Educational Testing Service.

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

Clark, B. R. (1983). *The higher education system: Academic organization in cross-national perspective*. Berkeley, CA: University of California Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

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

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

Davis, B. G. (2009). *Tools for teaching* (2nd ed.). San Francisco: Jossey-Bass.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annuals of Statistics, 7,* 1-26.

Efron, B., & Tibshirani, R. (1993a). Further topics in bootstrap confidence intervals *An introduction to the bootstrap* (pp. 321-337). Boca Raton, FL: Chapman & Hall.

Efron, B., & Tibshirani, R. (1993b). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall.

Erdle, S., & Murray, H. G. (1986). Interfaculty differences in classroom teaching behaviors and

their relationship to student instructional ratings. *Research in Higher Education, 24*(2), 115-127.

Erdle, S., Murray, H. G., & Rushton, J. P. (1985). Personality, classroom behavior, and student ratings of college teaching effectiveness: A path analysis. *Journal of Educational Psychology, 77*(4), 69-111.

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. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education, 10*(2), 149-172.

Feldman, K. A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education, 18*(1), 3-124.

Feldman, K. A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education, 21*(1), 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*(2), 129-213.

Feldman, K. A. (1989a). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30*(6), 583-645.

Feldman, K. A. (1989b). 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*(2), 137-194.

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*(3), 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*(2), 151-211.

Feldman, K. A. (1997). Identifying exemplary teachers and teaching. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 368-395). New York: Agathon Press.

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective*. Dordrecht, The Netherlands: Springer.

Franklin, J., & Theall, M. (1992). *Disciplinary differences: Instructional goals and activities, measures of student performance, and student ratings of instruction.* Paper presented at the Annual meeting of the American Educational Research Association, San Francisco.

Franklin, J., & Theall, M. (1995). The relationship of disciplinary differences and the value of class preparation time to student ratings of teaching. In N. Hativa & M. Marincovich (Eds.), *Disciplinary differences in teaching and learning* (pp. 41-48). San Francisco: Jossey-Bass.

Frey, P. W. (1976). Validity of student instructional ratings: Does timing matter? *Journal of Higher Education, 47*(3), 327-336.

Games, P. A., & Howell, J. F. (1976). Pairwise Multiple Comparison Procedures with Unequal

N's and/or Variances: A Monte Carlo Study. *Journal of Educational Statistics, 1*(2), 113-125.

Gilmore, G. (1994). *The effects of course demends and grading leniency on student ratings of instruction.* Paper presented at the American Educational Research Association, Atlanta, GA.

Holland, J. L. (1966). *The psychology of vocational choice*. Waltham, MA: Blaisdell.

Holland, J. L. (1973). *Making vocational choices: A theory of vocational personalities and work environments*. Engelwood Cliffs, NJ: Prentice-Hall.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments*. Odessa, FL: Psychological Assessment Resources.

Howard, G. S., & Maxwell, S. E. (1980). The correlation between student satisfaction and grades: A case of mistaken causation? *Journal of Educational Psychology, 72*(6), 810-820.

Howard, G. S., & Maxwell, S. E. (1982). Do grades contaminate student evaluations of instruction? *Research in Higher Education, 16*(2), 175-188.

Hoyt, D. P., & Cashin, W. E. (1977). IDEA technical report no. 1: Development of the IDEA system. Manhattan, KS: IDEA Center.

Hoyt, D. P., & Lee, E. (2002a). IDEA technical report no. 12: Basic data for the revised IDEA system. Manhattan, KS: IDEA Center.

Hoyt, D. P., & Lee, E. (2002b). IDEA technical report no. 13: Disciplinary differences in student ratings. Manhatten, KS: IDEA Center.

Kuh, G. D. (2001). Assessing what really matters to student learning: Inside the national survey of student engagement. *Change, 33*(3), 10-17, 66.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to Probability and Statistics* (pp. 278-292). Palo Alto, CA: Stanford University Press.

Marsh, H. W. (1982). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology, 74*(2), 264-279.

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

Marsh, H. W. (2001). Distinguishing between good (useful) and bad workloads on student evaluations of teaching. *American Educational Research Journal, 38*(1), 183-212.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319-383). Dordrecht, The Netherlands: Springer.

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 8, pp. 143-233). New York: Agathon Press.

Marsh, H. W., & Dunkin, M. J. (1997). Students evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241-320). New York: Agathon Press.

Marsh, H. W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching & Teacher Education, 7*(4), 303-314.

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Validity of student evaluations of

instructional effectiveness: A comparison of faculty self-evaluations and evaluation by their students. *Journal of Educational Psychology, 71*(2), 149-160.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, and innocent bystanders. *Journal of Educational Psychology*, *92*(1), 202-222.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Milem, J. F., & Umbach, P. D. (2003). Examining the perpetuation hypothesis: The influence of pre-college factors on students' predispositions regarding diversity activities in college. *Journal of College Student Development, 45*(5), 611-624.

Milem, J. F., Umbach, P. D., & Liang, C. (2004). Exploring the perpetuation hypothesis: The role of colleges and universities in desegregating society. *Journal of College Student Development, 45*(6), 688-700.

Morstain, B. R., & Smart, J. C. (1976). Educational orientations of faculty: Assessing a personality model of the academic professions. *Psychological Reports, 39,* 1199-1211.

Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology, 75*(1), 138-149.

Murray, H. G. (1985). Classroom teaching behaviors related to college teaching effectiveness. In J. G. Donald & A. M. Sullivan (Eds.), *Using Research to Improve Teaching*. San Francisco: Jossey Bass.

Murray, H. G., & Renaud, R. D. (1995). Disciplinary differerences in classroom teaching behaviors. In N. Hativa & M. Marincovich (Eds.), Disciplinary differences in teaching and learning: Implications for practice (pp. 31-39). *New Directions for Teaching and Learning*, 64. San Franscisco: Jossey-Bass.

Murray, H. G., Rushton, J. P., & Paunonen, S. V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology, 82*(2), 250-261.

Neumann, L., & Neumann, Y. (1985). Determimants of students' instructional evaluation: A comparision of four levels of academic areas. *Journal of Educational Research, 78*(3), 152-158.

Neumann, Y., & Neumann, L. (1983). Characteristics of academic areas and students' evaluation of instruction. *Research in Higher Education, 19*(3), 323-334.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? In M. Theall, P. C. Abrami & L. A. Mets (Eds.), The student ratings debate: Are they valid? How can we best use them? (pp. 27-43). *New Directions for Institutional Research,* 109. San Francisco, CA: Jossey-Bass.

Osborne, J. W. (2013). *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. Thousand Oaks, CA: SAGE.

Overall, J. U., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology, 72*(3), 321-325.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research*. San Franciso, CA: Jossey-Bass.

Peters, D. S. (1974). The link is equitability. *Research in Higher Education, 2*, 57-64.

Pohlmann, J. T. (1976). A description of effective college teaching in five disciplines as measured by student ratings. *Research in Higher Education, 4*, 335-346.

Remedios, R., & Lieberman, D. A. (2008). I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal, 34*(1), 91-115.

Renaud, R. D., & Murray, H. G. (1996). Aging, personality, and teaching effectiveness in academic psychologists. *Research in Higher Education, 37*(3), 323-340.

Resampling Stats for Excel (Version 4.0) [Computer software] Purchased and downloaded from http://www.resample.com/excel/

Roberts, M. J., & Russo, R. (1999). *A student's guide to analysis of variance.* New York, NY: Routledge.

Rosen, D., Holmberg, K., & Holland, J. L. (1997). *The educational opportunities finder*. Odessa, FL: Psychological Assessment Resources.

Ruscio, K. P. (1987). Many sectors, many professions. In B. R. Clark (Ed.), *The academic profession: National, disciplinary, and institutional settings*. Berkeley, CA: University of California Press.

Seldin, P. (1999). Current practices - good and bad - nationally. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions.* (pp. 1-24). Bolton, MA: Anker.

Sixbury, G. R., & Cashin, W. E. (1995). IDEA technical report no. 10: Comparative data by academic field. Manhattan: KS: IDEA Center.

Smart, J. C. (1982). Faculty teaching goals: A test of Holland's theory. *Journal of Educational*

Psychology, 74(2), 180-188.

Smart, J. C., Feldman, K. A., & Ethington, C. A. (2000). *Academic disciplines: Holland's theory and the study of college students and faculty*. Nashville: Vanderbilt University Press.

Smart, J. C., & Thompson, M. D. (2001). The environmental identity scale and the differentiation among environmental models in Holland's theory. *Journal of Vocational Behavior, 58*(3), 436-452.

Smart, J. C., & Umbach, P. D. (2007). Faculty and academic environments: Using Holland's theory to explore differences in how faculty structure undergraduate courses. *Journal of College Student Development, 48*(2), 183-195.

Smith, R. A., & Cranton, P. A. (1992). Students' perceptions of teachinig skills and overall effectiveness across instructional settings. *Research in Higher Education, 33*(6), 747-764.

Solomon, D. (1966). Teacher behavior dimensions, course characteristics, and student evaluations of teachers. *American Educational Research Journal, 3*, 35-47.

Spokane, A. R., Meir, E. I., & Catalano, M. (2000). Person-environment congruence and Holland's theory: A review and reconsideration. *Journal of Vocational Behavior, 57*(2), 137-187.

Theall, M., Franklin, J., & Ludlow, L. H. (1990). *Attributions and retributions: Student ratings and the percieved causes of performance.* Paper presented at the American Educational Research Association, Boston.

Thompson, M. D., & Smart, J. C. (1999). Student competencies emphasized by faculty in disparate environments. *Journal of College Student Development, 40*(4), 365-376.

Tsabari, T., Tziner, A., & Meir, E. I. (2005). Updated meta-analysis of the relationship between congruence and satisfaction. *Journal of Career Assessment, 13*(2), 216-232.

U.S. Department of Education. (1991). Assessing teaching performance. *The Department Chair: A Newsletter for Academic Administrators, 2*(3), 2.

Umbach, P. D. (2006). The contribution of faculty of color to undergraduate education. *Research in Higher Education, 47*(3), 317-345.

Umbach, P. D. (2007). Faculty cultures and college teaching. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 263-317). Dordrecht, The Netherlands: Springer.

Umbach, P. D., & Milem, J. F. (2004). Applying Holland's typology to the study of differences in student views about diversity. *Research in Higher Education, 45*(6), 625-649.

Umbach, P. D., & Wawrzynski, M. R. (2005). Faculty do matter: The role of college faculty in student learning and engagement. *Research in Higher Education, 46*(2), 153-184.

Wilcox, R. R. (1998). A note on the Theil-Sen regression estimatior when the regressor is random and ther eerror term is heteroscedastic. *Biometrical Journal, 40*(3), 261-268.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. Amsterdam; Boston: Elsevier/Academic Press.

Wright, D. B., London, K., & Field, A. P. (2011). Using bootstrap estimation and the plug-in principle for clinical psychology data. *Journal of Experimental Psychopathology, 2*(2), 252-270.

Appendix A

IRB Approval

VirginiaTech

Office of Research Compliance
Institutational Review Board
North End Center, Suite 4120, Virginia Tech
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-4606 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE: June 25, 2013

TO: Anne Margaret Laughlin, Yasuo Miyazaki, Steven M Janosik

FROM: Virginia Tech Institutional Review Board (FWA00000572, expires April 25, 2018)

PROTOCOL TITLE: Student Perceptions of Teaching: Examining Contexts for Interpretation

IRB NUMBER: **12-632**

Effective June 25, 2013, the Virginia Tech Institution Review Board (IRB) Chair, David M Moore, approved the Continuing Review request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As:	**Expedited, under 45 CFR 46.110 category(ies) 5**
Protocol Approval Date:	**July 10, 2013**
Protocol Expiration Date:	**July 9, 2014**
Continuing Review Due Date*:	**June 25, 2014**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

Appendix B

Class Subjects Categorized by Academic College and Holland Environment

Sorted by Environment within College

Academic College	Class Subject	Holland Environment
Liberal Arts & Human Sciences	English	Artistic
Liberal Arts & Human Sciences	Fine Arts	Artistic
Liberal Arts & Human Sciences	Foreign Languages	Artistic
Liberal Arts & Human Sciences	Music	Artistic
Liberal Arts & Human Sciences	Philosophy	Artistic
Liberal Arts & Human Sciences	Theatre Arts	Artistic
Liberal Arts & Human Sciences	Education	Social
Liberal Arts & Human Sciences	Human Development	Social
Liberal Arts & Human Sciences	Interdisciplinary Studies - Humanities	Social
Liberal Arts & Human Sciences	Interdisciplinary Studies - Religion	Social
Liberal Arts & Human Sciences	Interdisciplinary Studies - Science and Technology	Social
Liberal Arts & Human Sciences	Science and Technology Studies	Social
Liberal Arts & Human Sciences	Military Aerospace Studies	Realistic
Liberal Arts & Human Sciences	Military Navy	Realistic
Liberal Arts & Human Sciences	Military Science (AROTC)	Realistic
Liberal Arts & Human Sciences	Apparel, Housing, and Resource	Enterprising
Liberal Arts & Human Sciences	Communication Studies	Enterprising
Liberal Arts & Human Sciences	History	Enterprising
Liberal Arts & Human Sciences	Interdisciplinary Studies - Communication Studies	Enterprising
Liberal Arts & Human Sciences	Interdisciplinary Studies - History	Enterprising
Liberal Arts & Human Sciences	Interdisciplinary Studies - Political Science	Enterprising
Liberal Arts & Human Sciences	International Studies	Enterprising
Liberal Arts & Human Sciences	Political Science	Enterprising

Liberal Arts & Human Sciences	Interdisciplinary Studies - Agricultural and Applied Economics	Investigative
Liberal Arts & Human Sciences	Interdisciplinary Studies - Geography	Investigative
Liberal Arts & Human Sciences	Sociology	Investigative
Liberal Arts & Human Sciences	Liberal Arts and Human Sciences - General	Not Classified
Liberal Arts & Human Sciences	Interdisciplinary Studies	Not Classified
Agriculture & Life Sciences	Agricultural and Extension Education	Social
Agriculture & Life Sciences	Human Nutrition, Foods and Exercise	Social
Agriculture & Life Sciences	Animal and Poultry Sciences	Realistic
Agriculture & Life Sciences	Dairy Science	Realistic
Agriculture & Life Sciences	Horticulture	Realistic
Agriculture & Life Sciences	Agricultural and Applied Economics	Investigative
Agriculture & Life Sciences	Agriculture and Life Sciences	Investigative
Agriculture & Life Sciences	Biochemistry	Investigative
Agriculture & Life Sciences	Crop and Soil Environmental Sciences	Investigative
Agriculture & Life Sciences	Entomology	Investigative
Agriculture & Life Sciences	Food Science and Technology	Investigative
Agriculture & Life Sciences	Plant Pathology, Physiology and Weed	Investigative
Architecture	Architecture	Artistic
Architecture	Art and Art History	Artistic
Architecture	Interior Design	Artistic
Architecture	Building Construction	Realistic
Architecture	School of Public and International Affairs	Enterprising
Business	Finance	Enterprising
Business	Hospitality and Tourism Management	Enterprising
Business	Management	Enterprising
Business	Marketing	Enterprising
Business	Accounting and Information Systems	Investigative
Business	Business Information Technology	Investigative
Science	Biological Sciences	Investigative
Science	Chemistry	Investigative

Science	Economics	Investigative
Science	Geosciences	Investigative
Science	Mathematics	Investigative
Science	Physics	Investigative
Science	Psychology	Investigative
Science	Statistics	Investigative
Science	College Of Science - General	Not Classified
Engineering	Materials Science and Engineering	Realistic
Engineering	Mechanical Engineering	Realistic
Engineering	Mining Engineering	Realistic
Engineering	Myers-Lawson School of Construction	Realistic
Engineering	Aerospace and Ocean Engineering	Investigative
Engineering	Biological Systems Engineering	Investigative
Engineering	Biomedical Engineering and Sciences	Investigative
Engineering	Chemical Engineering	Investigative
Engineering	Civil and Environmental Engineering	Investigative
Engineering	Computer Science	Investigative
Engineering	Electrical and Computer Engineering	Investigative
Engineering	Engineering Education	Investigative
Engineering	Engineering Science and Mechanics	Investigative
Engineering	Industrial and Systems Engineering	Investigative
Engineering	Engineering - General	Not Classified

Sorted by College within Environment

Holland Environment	Class Subject	Academic College
Artistic	English	Liberal Arts & Human Sciences
Artistic	Fine Arts	Liberal Arts & Human Sciences
Artistic	Foreign Languages	Liberal Arts & Human Sciences
Artistic	Music	Liberal Arts & Human Sciences
Artistic	Philosophy	Liberal Arts & Human Sciences
Artistic	Theatre Arts	Liberal Arts & Human Sciences
Artistic	Architecture	Architecture
Artistic	Art and Art History	Architecture
Artistic	Interior Design	Architecture
Social	Education	Liberal Arts & Human Sciences
Social	Human Development	Liberal Arts & Human Sciences
Social	Interdisciplinary Studies - Humanities	Liberal Arts & Human Sciences
Social	Interdisciplinary Studies - Religion	Liberal Arts & Human Sciences
Social	Interdisciplinary Studies - Science and Technology	Liberal Arts & Human Sciences
Social	Science and Technology Studies	Liberal Arts & Human Sciences
Social	Agricultural and Extension Education	Agriculture & Life Sciences
Social	Human Nutrition, Foods and Exercise	Agriculture & Life Sciences
Realistic	Military Aerospace Studies	Liberal Arts & Human Sciences
Realistic	Military Navy	Liberal Arts & Human Sciences
Realistic	Military Science (AROTC)	Liberal Arts & Human Sciences
Realistic	Animal and Poultry Sciences	Agriculture & Life Sciences
Realistic	Dairy Science	Agriculture & Life Sciences
Realistic	Horticulture	Agriculture & Life Sciences
Realistic	Building Construction	Architecture
Realistic	Materials Science and Engineering	Engineering
Realistic	Mechanical Engineering	Engineering
Realistic	Mining Engineering	Engineering
Realistic	Myers-Lawson School of Construction	Engineering

Enterprising	Apparel, Housing, and Resource	Liberal Arts & Human Sciences
Enterprising	Communication Studies	Liberal Arts & Human Sciences
Enterprising	History	Liberal Arts & Human Sciences
Enterprising	Interdisciplinary Studies - Communication Studies	Liberal Arts & Human Sciences
Enterprising	Interdisciplinary Studies - History	Liberal Arts & Human Sciences
Enterprising	Interdisciplinary Studies - Political Science	Liberal Arts & Human Sciences
Enterprising	International Studies	Liberal Arts & Human Sciences
Enterprising	Political Science	Liberal Arts & Human Sciences
Enterprising	School of Public and International Affairs	Architecture
Enterprising	Finance	Business
Enterprising	Hospitality and Tourism Management	Business
Enterprising	Management	Business
Enterprising	Marketing	Business
Investigative	Interdisciplinary Studies - Agricultural and Applied Economics	Liberal Arts & Human Sciences
Investigative	Interdisciplinary Studies - Geography	Liberal Arts & Human Sciences
Investigative	Sociology	Liberal Arts & Human Sciences
Investigative	Agricultural and Applied Economics	Agriculture & Life Sciences
Investigative	Agriculture and Life Sciences	Agriculture & Life Sciences
Investigative	Biochemistry	Agriculture & Life Sciences
Investigative	Crop and Soil Environmental Sciences	Agriculture & Life Sciences
Investigative	Entomology	Agriculture & Life Sciences
Investigative	Food Science and Technology	Agriculture & Life Sciences
Investigative	Plant Pathology, Physiology and Weed	Agriculture & Life Sciences
Investigative	Accounting and Information Systems	Business
Investigative	Business Information Technology	Business
Investigative	Biological Sciences	Science
Investigative	Chemistry	Science
Investigative	Economics	Science
Investigative	Geosciences	Science
Investigative	Mathematics	Science

Investigative	Physics	Science
Investigative	Psychology	Science
Investigative	Statistics	Science
Investigative	Aerospace and Ocean Engineering	Engineering
Investigative	Biological Systems Engineering	Engineering
Investigative	Biomedical Engineering and Sciences	Engineering
Investigative	Chemical Engineering	Engineering
Investigative	Civil and Environmental Engineering	Engineering
Investigative	Computer Science	Engineering
Investigative	Electrical and Computer Engineering	Engineering
Investigative	Engineering Education	Engineering
Investigative	Engineering Science and Mechanics	Engineering
Investigative	Industrial and Systems Engineering	Engineering
Not Classified	Interdisciplinary Studies	Liberal Arts & Human Sciences
Not Classified	Liberal Arts and Human Sciences - General	Liberal Arts & Human Sciences
Not Classified	College Of Science - General	Science
Not Classified	Engineering - General	Engineering

Appendix C

Minimum Detectable Effect Sizes

Statistical test	Sample[a]	df	Minimum detectable effect size (*f*)[b]
Main effect of academic field	6 groups x 204 classes per group	5	.10
Main effect of class size and main effect of course level	2 groups x 1667 classes per group	1	.05
Three-way interaction effect for academic field, class size, and course level	24 groups x 21 classes per group	5	.16
Two-way interaction effect for class size and course level	4 groups x 21 classes per group	1	.31

Note. Calculations are based on alpha = .05 and power = .80.

[a]The number of classes are based on the smallest group in the test design; this provides the most conservative estimates. [b]Cohen's (1988, pp. 285 – 288) rules of thumb for effect size: $f = .10$ is small, $f = .25$ is medium, $f = .40$ is large.

Appendix D

Bootstrap Procedure for Research Question Seven

For research question seven, bootstrapping was used to conduct a hypothesis test in a situation where no standard statistical test exists. This research question involved comparing the eta squared from the ANOVA for academic college ($\eta^2_{college}$) with the eta squared from the ANOVA for Holland environment ($\eta^2_{Holland}$). To determine the statistical significance of a difference between the two eta squared values ($\eta^2_{college} - \eta^2_{Holland}$), a bootstrap procedure was used to estimate a sampling distribution for the difference. This analysis was done using Resampling Stats for Excel (Version 4.0), an add-in for Microsoft Excel that facilitates bootstrapping and permutation procedures[12]. The resampling was done within cells of a two-way table that crossed the two classification methods. Table D1 shows the number of ratings in each cell in the two-way classification.

In the bootstrap, the original data were treated as a population, and 10,000 bootstrap samples, each of which contained *N* (= 4,599) cases, were resampled with replacement. The original data were organized in a spreadsheet so that the data from each cell in the two-way table (Table D1) were represented in a separate column.

Table D2 shows the layout of the original data. Each cell in this spreadsheet contains the mean rating for a particular class in a particular college (indicated in row 2). Each class was also categorized into a particular Holland environment (indicated in row 3). For example cell A:4 contains the mean rating (3.19) for a class in the College of Agriculture and Life Sciences (CALS) and Holland's investigative environment (Inv). Cell E:4 contains the mean rating (3.06)

[12] Resampling Stats for Excel (Version 4.0) [Computer software] Purchased and downloaded from http://www.resample.com/excel/

Table D1

Crosstab of Six Academic Colleges by Five Holland Environments: Number of Class Ratings in Each Group

	Holland Environment					
Academic College	Artistic	Social	Realistic	Enterprising	Investigative	Total
Liberal Arts & Human Sciences	734	146	56	397	89	1,422
Agriculture & Life Sciences	0	101	97	0	154	352
Architecture	151	0	27	26	0	204
Business	0	0	0	246	180	426
Science	0	0	0	0	1,394	1,394
Engineering	0	0	157	0	644	801
Total	885	247	337	669	2,461	4,599

Table D2

Layout of Original Data with a Column for Each Group based on Academic College and Holland Environment.

ORIGINAL DATA																			
CALS			CAUS			CLAHS					COB		COE		COS				
Inv	Real	Soc	Art	Ent	Real	Art	Ent	Inv	Real	Soc	Ent	Inv	Inv	Real	Inv				
3.19	4.11	2.36	3.06	2.73	3.47	2.10	2.95	3.08	4.56	2.58	2.65	1.78	2.14	2.40	1.76				
3.24	4.13	2.90	3.33	3.20	3.58	3.00	3.00	3.85	4.85	2.75	3.03	2.30	2.70	2.63	1.83				
3.31	4.59	3.00	3.38	3.38	3.83	3.08	3.13	3.91	4.86	2.86	3.31	2.43	2.73	2.67	1.86				
3.75	4.59	3.33	3.55	3.95	3.97	3.33	3.17	4.06	4.94	3.41	3.33	2.53	2.75	2.81	1.86				
3.82	4.64	3.69	3.73	3.96	4.05	3.42	3.25	4.24	5.00	3.57	3.38	2.78	2.80	2.91	2.06				
3.94	4.68	3.81	3.77	4.10	4.09	3.56	3.29	4.36	5.09	3.69	3.39	3.00	2.91	3.17	2.19				
4.04	4.69	4.00	4.01	4.33	4.18	3.64	3.40	4.45	5.13	4.04	3.58	3.05	2.98	3.22	2.30				
4.05	4.70	4.18	4.02	4.34	4.42	3.70	3.45	4.50	5.18	4.09	3.61	3.05	3.00	3.31	2.32				
4.09	4.73	4.20	4.06	4.38	4.68	3.77	3.50	4.52	5.19	4.12	3.64	3.06	3.05	3.45	2.32				
4.19	4.73	4.33	4.15	4.60	4.76	3.82	3.50	4.52	5.25	4.14	3.65	3.10	3.06	3.47	2.35				
4.20	4.83	4.38	4.24	4.68	4.85	3.82	3.52	4.53	5.29	4.15	3.65	3.18	3.09	3.53	2.39				
4.20	4.88	4.40	4.38	4.71	4.90	3.88	3.60	4.57	5.30	4.17	3.70	3.29	3.09	3.56	2.44				
4.29	4.92	4.41	4.44	4.77	5.00	3.90	3.62	4.64	5.33	4.23	3.73	3.33	3.11	3.65	2.48				
4.30	4.93	4.53	4.50	4.84	5.27	3.90	3.64	4.67	5.33	4.24	3.83	3.43	3.15	3.67	2.50				
4.33	4.98	4.59	4.50	4.87	5.27	3.93	3.67	4.68	5.40	4.28	3.90	3.44	3.18	3.68	2.52				
4.34	5.00	4.64	4.50	4.90	5.35	3.93	3.67	4.73	5.46	4.30	3.93	3.49	3.25	3.70	2.53				
4.38	5.00	4.74	4.56	4.92	5.50	3.94	3.75	4.81	5.50	4.30	3.95	3.59	3.28	3.75	2.62				
4.38	5.03	4.77	4.60	4.96	5.50	4.00	3.77	4.82	5.50	4.33	3.98	3.69	3.29	3.86	2.66				
4.40	5.06	4.77	4.60	5.14	5.50	4.07	3.79	4.85	5.50	4.33	4.00	3.71	3.29	3.94	2.67				
4.45	5.10	4.79	4.64	5.18	5.50	4.09	3.80	4.90	5.50	4.36	4.00	3.81	3.40	3.97	2.71				
4.45	5.11	4.83	4.68	5.31	5.56	4.12	3.83	4.90	5.55	4.40	4.00	3.87	3.41	4.00	2.72				
4.45	5.13	4.86	4.68	5.40	5.61	4.13	3.83	4.92	5.56	4.44	4.06	4.00	3.43	4.00	2.73				
4.47	5.14	4.86	4.73	5.47	5.64	4.13	3.91	4.96	5.58	4.47	4.07	4.00	3.44	4.02	2.74				
4.47	5.15	4.92	4.73	5.50	5.69	4.15	4.00	5.00	5.60	4.47	4.08	4.01	3.44	4.03	2.76				
4.47	5.15	4.93	4.74	5.63	5.70	4.15	4.00	5.00	5.60	4.50	4.14	4.07	3.45	4.06	2.77				
4.54	5.17	4.95	4.75	5.66	5.95	4.18	4.00	5.03	5.63	4.50	4.15	4.13	3.46	4.07	2.78				
4.54	5.17	5.00	4.80		6.00	4.20	4.05	5.05	5.63	4.53	4.17	4.21	3.47	4.13	2.80				
4.63	5.22	5.05	4.83			4.21	4.05	5.06	5.64	4.58	4.23	4.29	3.48	4.15	2.80				
4.65	5.24	5.07	4.83			4.21	4.05	5.08	5.68	4.63	4.27	4.29	3.49	4.15	2.86				
4.69	5.26	5.10	4.83			4.21	4.08	5.09	5.69	4.63	4.27	4.36	3.50	4.17	2.87				
4.75	5.29	5.10	4.83			4.23	4.11	5.12	5.70	4.65	4.27	4.37	3.54	4.19	2.90				

for a class in the College of Architecture and Urban Studies (CAUS) and Holland's artistic environment (Art).

Eta squared was calculated for each classification method by conducting one-way ANOVAs on the resampled data. In Table D3, rows 23 to 51 show some of the data from one of the bootstrap samples. Rows 4 to 8 show results from the two one-way ANOVAs that were calculated based on this bootstrap sample. The ANOVA table on the left is based on the academic college classification and the ANOVA table on the right is based on the Holland classification. The eta squared value for each classification method is highlighted in pink.

In Table D3, row 13 shows the sample size for each group. The values show that group sizes in the bootstrap sample matched group sizes in the original sample for both the academic college classification (CALS = 352, CAUS = 204, CLAHS = 1422, COB = 426, COE = 801, COS = 1394) and the Holland environment classification (Artistic = 885, Enterprising = 669, Investigative = 2461, Realistic = 337, Social = 247).

At the end of each iteration of bootstrapping, the eta squared value for each classification method was computed and saved along with the eta squared difference. This process produced a list of 10,000 values for each classification method, and for the eta squared difference. The 10,000 values for the difference provided an estimate of the sampling distribution of the eta squared difference (see Figure 18 in Chapter 4). By sorting these values and finding the 2.5^{th} and 97.5^{th} percentiles, the bootstrap 95% confidence interval for the eta squared difference was estimated. These results are shown in Table D4.

Table D3

Eta Squared Values for Each Classification Method Based on a Bootstrap Sample

ANALYSIS BASED ON BOOTSTRAP SAMPLE

ANOVA for Holland Environment

	SS	*df*	*MS*	*F*	eta^2
Between	111.42	4	27.86	65.946	0.054
Within	1940.54	4594	0.42		
Total	2051.97	4598			

grand mean: 5.031

	Art	**Ent**	**Inv**	**Rea**	**Soc**
n	885	669	2461	337	247
mean	5.30	5.10	4.90	5.09	5.13
between	62.29	2.83	42.56	1.24	2.51
within	240.36	259.91	1181.67	168.58	90.03

ANOVA for Academic College

	SS	*df*	*MS*	*F*	eta^2
Between	136.79	5	27.36	65.608	0.067
Within	1915.18	4593	0.42		
Total	2051.97	4598			

grand mean: 5.031

	CALS	**CAUS**	**CLAHS**	**COB**	**COE**	**COS**
n	352	204	1422	426	801	1394
mean	5.21	5.09	5.24	4.99	5.09	4.87
between	11.44	0.77	64.11	0.62	3.02	34.07
within	101.09	102.44	420.29	210.62	414.66	666.09

BOOTSTRAP SAMPLE

CALS			**CAUS**			**CLAHS**					**COB**	
Inv	**Real**	**Soc**	**Art**	**Ent**	**Real**	**Art**	**Ent**	**Inv**	**Real**	**Soc**	**Ent**	**Inv**
4.54	5.53	5.21	5.53	4.60	4.85	5.82	5.72	5.30	5.63	5.18	3.93	3.59
4.05	5.69	5.11	5.14	4.90	3.58	5.70	5.18	5.30	5.89	5.07	5.20	5.38
4.45	4.98	5.40	5.31	4.38	5.00	5.50	5.50	5.49	5.80	5.27	6.00	5.83
4.95	4.70	4.74	5.32	2.73	5.64	5.86	5.38	5.18	5.29	4.33	5.28	5.12
4.89	4.69	5.69	5.87	5.14	5.61	4.64	5.45	5.38	5.80	3.41	5.31	5.59
5.12	5.68	5.92	5.68	4.60	5.50	4.85	4.67	4.45	5.50	4.50	5.16	4.54
5.65	5.59	5.71	5.64	5.50	4.76	5.77	4.18	5.46	5.81	5.74	5.20	5.32
5.30	5.81	5.73	5.67	3.95	4.76	5.90	5.48	5.82	5.33	5.58	4.27	5.48
4.54	5.77	4.77	5.47	4.92	3.83	5.50	5.70	5.67	5.64	4.15	4.33	4.80
4.92	5.50	5.67	4.50	4.34	3.97	4.42	5.86	4.50	5.76	4.33	5.47	3.49
5.50	5.22	5.61	5.80	4.10	5.50	5.08	4.71	5.25	5.50	5.10	5.73	5.31
4.75	5.06	4.18	4.44	4.10	5.69	5.45	5.55	5.54	4.85	5.25	3.58	5.37
5.65	5.81	5.76	5.70	4.34	5.35	5.27	5.93	4.24	5.60	5.40	4.92	4.50
5.35	5.11	5.10	5.36	2.73	5.35	6.00	5.53	5.38	5.60	5.53	5.21	4.57
5.60	4.11	4.00	5.43	5.63	5.64	4.92	4.69	5.67	4.56	5.50	5.67	4.46
5.25	5.50	3.81	5.64	3.20	5.50	5.19	3.29	5.09	5.70	4.82	5.41	3.81
5.00	5.81	5.54	5.79	3.95	5.64	4.59	5.60	5.86	4.94	5.00	4.60	5.55
5.30	5.15	5.57	5.62	5.63	4.42	5.53	5.80	5.65	5.91	4.82	5.05	5.17
5.43	5.96	5.60	5.60	4.60	5.70	5.18	4.98	5.31	5.82	5.94	2.65	5.45
4.99	5.81	5.82	5.78	5.18	3.97	4.86	5.34	5.24	5.13	3.41	5.43	4.57
4.40	4.69	4.20	5.64	3.96	3.83	5.82	5.69	5.30	5.29	5.81	5.18	5.76
5.15	4.98	6.00	5.82	4.90	3.47	5.46	3.17	5.50	5.88	4.14	5.78	4.59
4.77	5.77	6.00	5.77	4.84	4.18	5.55	5.33	5.30	5.30	5.00	5.00	3.69
4.89	4.68	4.64	4.75	4.10	5.95	5.72	4.18	5.44	5.19	5.81	4.71	3.59
4.89	5.72	5.76	4.75	5.14	5.50	5.33	4.94	4.64	5.82	4.97	5.33	5.42
4.95	5.06	4.83	5.67	3.20	5.64	5.11	5.75	5.36	5.73	5.94	5.13	5.44
5.05	5.42	4.77	4.95		5.27	4.42	5.67	4.53	5.70	4.94	5.44	3.29
5.21	5.50	5.50	5.80			5.44	5.67	5.00	5.92	5.19	5.70	4.01
5.17	5.68	4.77	6.00			6.00	5.80	5.65	5.71	5.00	5.25	4.00

Table D4

Result Sheet for Bootstrap Procedure

	A	B	C	D	E	F
1	**Eta Squared Environment**	**Eta Squared College**	**Eta Squared Difference**			
2	0.048	0.059	0.011			
3	0.045	0.057	0.012			
4	0.062	0.077	0.015			
5	0.050	0.062	0.012			
6	0.050	0.068	0.017		**2.5 percentile of difference:**	**0.004**
7	0.060	0.066	0.005			
8	0.055	0.071	0.016		**97.5 percentile of difference:**	**0.022**
9	0.070	0.081	0.011			
10	0.056	0.061	0.005		**Mean of difference:**	**0.013**
11	0.057	0.075	0.019			
12	0.059	0.075	0.016		**Standard error of difference:**	**0.005**
13	0.057	0.073	0.016			
14	0.055	0.075	0.020			
15	0.050	0.062	0.012			
16	0.059	0.082	0.023			
17	0.065	0.079	0.014			
18	0.066	0.073	0.007			
19	0.046	0.069	0.022			
20	0.064	0.075	0.011			
21	0.054	0.066	0.012			
22	0.058	0.073	0.015			
23	0.050	0.060	0.010			
24	0.052	0.063	0.011			
25	0.050	0.060	0.010			