EXPLORING SET INSTRUCTOR, COURSE, AND STUDENT BIASES IN A LARGE, URBAN, PUBLIC, R1 BUSINESS SCHOOL

A Dissertation
Submitted to the
Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

---

by
Matthew Kunkle

Dissertation Sponsoring Committee Members (June, 2020):
Dr. Joseph DuCette, Policy, Organizational and Leadership Studies, College of Education
Dr. Gary Blau, Human Resource Management, Fox School of Business
Dr. Joseph Paris, Policy, Organizational and Leadership Studies, College of Education
Outside Reader:
Dr. Jodi Levine-Laufgraben, Vice Provost for Academic Affairs, Assessment and Institutional Research

ABSTRACT

Using the SET questions as the dependent variable(s), this study answers the following research questions: (1) Are the Instructor variables of gender and race biasing factors?; (2) Are the Course variables of class size and content (qualitative to quantitative ratio) biasing factors?; (3) Is the Student variable of section-level GPA related to SET ratings?; (4) Is the administrative mode of data collection, the change from pencil and paper to online data collection, a biasing factor? Questions were answered through bivariate correlations and two-way repeated measures ANOVAs. This study found there was a significant effect on SET outcomes as a function of race, but not for gender. While class size had no significant effect on SET outcomes, the section-level GPA and the amount of qualitative vs quantitative course content did. The administrative mode of data collection had a significant effect, mostly due to the large sample size.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND BACKGROUND

In both business and education, 360 degree feedback is vital to success, and employees need to be able to take constructive feedback. In my background as a musician, entrepreneur, technical support supervisor, educator, marching band leader, and corporate compliance manager, I found that all of these roles required harmony of shared goals and proper communication. As an educator, the integrity of impartial evaluation weighed heavily on me. Is it based on a growth model or a flat plane? How do I keep my biases in check? As it turns out, actionable feedback from students, and some self-awareness and self reflection, can greatly improve my teaching effectiveness.

Student Evaluations of Teaching (SETs), also called "course evaluations," have been used in higher education for more than 10 decades (Brandenburg & Remmers, 1927). As Spooren, Brockx and Mortelmans (2013) point out, these assessments have three broad goals: to improve teaching quality; to provide input for appraisal purposes (e.g., tenure and promotion decisions, merit allocation); to provide evidence for institutional accountability.

In recent years, there has been a growing body of criticisms of these evaluations (McPherson, 2006; Uttl, Carmela, White, & Gonzalez, 2017). In these criticisms of SETs there is very little indication that anyone is concerned about the first reason: improving teaching quality. That is, student evaluations of teaching used by individual instructors to assess, reflect, and then improve their teaching are a valuable and unchallenged method that all instructors should use. Saying this another way, when SETs are used formatively, there is almost no criticism about their use. Moreover, giving students the ability to

provide feedback to instructors signals that the institution is concerned about the quality of its teaching, and that it cares about what its students think.

The concerns are raised about the other two reasons which are used to provide evidence for institutional accountability, especially input for appraisal purposes. It is when SETs are used summatively, especially if they are used as the sole or major source of data, that the problems arise; such has been commonplace since the 1970s (McKeachie, 1979). There are quite a few issues that have been raised about SETs – for example, validity and reliability concerns – but the major focus is on bias. The literature on bias can be organized into two broad categories: factors relating to the instructor and factors relating to the course. Among the factors listed for the instructor are gender, race, language background, attractiveness, rank, grading practices and enthusiasm. Course factors include class size, content, level, required versus elective, General Education versus specific education and workload. A summary of these factors mentioned in the literature, and specifically by Spooren et al., are presented in Tables 1.1 and 1.2.

Table 1.1

Instructor Characteristics

Characteristic	Nature of the Concern	Findings
Instructor gender	Female instructors receive lower SETs than male instructors.	Despite what is commonly mentioned in articles, the evidence about gender bias is actually mixed. In fact, there are more articles that show that female instructors receive higher SETs than males.
Instructor race	Minority instructors receive lower SETs than White instructors.	Here the evidence is more consistent: minority instructors in general, and African American instructors in particular, receive lower SETs.
Instructor language background	Instructors for whom English is a second language receive lower SETs	There is rather consistent evidence that this concern is warranted, especially in science.
Instructor attractiveness	Good looking instructors receive higher SETs.	The evidence supports this: likable, good-looking and well-dressed instructors receive higher SETs.
Instructor Rank	Full professors receive lower SETs.	The evidence for this is also mixed. The most consistent finding is that TAs typically receive higher SETs.
Instructor grading policies	Expected grade correlates positively with SETs—instructors buy SETs with high grades.	As mentioned above, this is a consistent, although controversial, finding. Of all the factors listed expected grade typically has the largest effect.
Instructor enthusiasm	Enthusiastic instructors get higher SETs (the Dr. Fox effect)	An often-cited experiment supports this.

Table 1.2

Course Characteristics

Characteristic	Nature of the Concern	Findings
Class size	Larger classes receive lower SETs.	This is generally supported although there is confounding with course content (e.g., large introductory science courses)
Course discipline	Natural science courses receive lower SETs.	Almost all of the research supports this.
Course level	Lower level courses receive lower SETs.	This is generally supported, although there are some confounding factors (see below).
Required vs. Elective	Required courses receive lower SETs.	This is generally supported.
General Education vs. specific education	General Education courses receive lower SETs.	Almost all of the research supports this.
Course workload	Courses that require more work receive lower SETs.	The evidence here is actually more contradictory than supportive. The most consistent finding is that courses that are viewed as "just right" in terms of workload are rated higher. Other than that, courses that require more work are actually rated higher.

As the authors point out, however, some of the factors that have been shown to correlate with SETs are not actually evidence of bias. As they say:

> Some of the factors are meaningful indicators of student learning and are therefore logically related to effective teaching and SET. For example, student effort and class attendance indicate the interest and motivation of students in a particular course and are at least partly dependent upon the organization of and the teaching in that course. The experience and research productivity of the teacher are valuable indicators of a teacher's education skills and knowledge of the subject matter. (p. 609)

The authors also point out that there are issues raised in the bias literature that are ambiguous as to their meaning, such as students' expected grade. It is well established that there is a positive correlation between the grade a student expects to get in the course and the student's evaluation of the course and the instructor (Stroebe, 2016). Those who

view this as bias say that this correlation shows that instructors can "buy" good evaluations by giving high grades, and cite this fact as one of the major reasons for grade inflation. The proponents of SETs say that this correlation simply reflects the fact that students who learn more in a course rate the instructor higher because they believe that the instructor is one of the reasons they obtained the higher grade. Although both arguments are reasonable, a recent meta-analysis (Uttl, White, & Gonzalez, 2017) shows that there is almost no relationship between SET's and knowledge gained in the course. This strengthens the argument that SET's are one of the reasons for the problem with grade inflation.

The non-ambiguous issue of bias is raised about teacher or course characteristics that have nothing to do with teaching quality. Among these are the instructor's gender, race and sexual orientation. If it is the case that instructors in specific groups (for example, African American females) obtain lower SETs simply because of their race and/or gender, and if this effect is significant, then it is argued that the SETs are so flawed that they shouldn't be used.

There is a wave of activity from institutions looking for a more holistic and balanced way to measure teaching effectiveness. In 2018, the University of Southern California (USC) announced SETs would no longer be an element of tenure and promotion review, but would still be used as a faculty reflective instrument to improve their instructional design (Carter, 2018). After a correction in 2018, USC currently uses a mainly peer-review process, with student evaluations playing only a part in measuring student engagement (Doerer, 2019). In 2019, the University of Oregon developed a new, holistic teaching evaluation system that does more than merely replace problematic SETs,

but puts them in balance with faculty self-reflection and peer evaluations (Provost, 2019). The following are examples of institutions that are currently seeking alternative methods of evaluating teaching effectiveness: Colorado State University at Fort Collins, the University of Colorado at Boulder, the University of Kansas, the University of Massachusetts at Amherst, Ryerson University, in Toronto, and a division of the University of California at Berkeley (Doerer, 2019).

While research on SETs is abundant, gaps in the literature still remain. Table 3 details the research questions of this study, and aligns them with gaps in the literature.

Table 1.3

Research Questions and Gaps in the Literature

Research Questions using SETs as the DV	Gaps in the Literature
Are the Instructor variables of gender and race biasing factors?	The SET literature typically addresses trends across the entire university. There hasn't been much attention in the literature exploring whether these biasing factors are true in specific contexts such as business schools where they have their own cultural norms and biases.
Are the Course variables of class size and content (qualitative to quantitative ratio) biasing factors?	Class size has either had a negative or no effect on SETs. The issue of course content is broadly defined, usually based on assumptions. Courses are typically designated quantitative or qualitative, without "a little of both" option, and this was presumed by department or by program. For example, a course on the ethics of accounting would be labeled quantitative because it is housed in the accounting department, but the course could be all written communication taught by a lawyer.
What is the effect of moving from pencil and paper to online SET forms?	Some studies have been conducted on response rates but nobody has systematically looked at whether moving from paper administration to online has additional effects. This study explored whether SET rating are affected by moving from pencil and paper to online SET forms.

The purpose of the present study was to investigate whether the above factors (instructor variables, course variables, SET administration modality) shown to be biasing in the literature, are also significant in a business school housed within a large, urban,

public, R1[1] institution. Data in this study include all relevant undergraduate business school sections spanning six years from fall 2009 through spring 2015. The variables in these data include: Instructor variables of gender and race; Course variables of class size and the ratio of qualitative to quantitative content; Student variables of section-level student GPA data; and SET data collection method (paper versus digital).

The institution's SET form data, both paper SET and electronic SET, are reported at the course section-level. In addition, because the institution went from collecting SET data though a traditional pencil and paper method to an eSET online collection method in fall 2012 - the mid-point of my dataset - I also investigated whether the above biasing factors differ between traditional pencil and paper and online collection methods.

Using the SET questions as the dependent variable(s), research questions are:

(1) Are the Instructor variables of gender and race biasing factors?

(2) Are the Course variables of class size and content (qualitative to quantitative ratio) biasing factors?

(3) Is the Student variable of section-level GPA related to SET ratings?

(4) Is the data collection method, the change from pencil and paper to online data collection, a biasing factor?

[1] R1=Research 1 Doctoral Universities (Very high research activity). This is the highest Carnegie research activity classification.

CHAPTER 2

REVIEW OF THE LITERATURE

Although not perfect, there are strong arguments that support the contention that SETs provide valuable feedback about an instructor's teaching effectiveness (McKeachie & Svinicki, 2006). Some authors argue that SETs are reliable as both a direct measure of student satisfaction with instruction and as an indirect measure of student learning (e.g., Marsh, 2007; Murray, 2007). Quantitative evaluations of instructors' overall teaching effectiveness, using solely SETs are frequently emphasized in personnel decisions (Centra & Gaubatz, 2000; Chan, Luk, & Zeng, 2014). At the institution in this study, along with other Higher Education institutions (Abrami, d'Apollonia, & Rosenfield, 2007), SETs also play an important role in the selection of teaching award winners, institutional program reviews, and student course selection. In fact, selected aggregate SET outcomes are shared with and used as an incentive to increase response rates.

Of great importance to the careers of instructors, these ratings are "used by faculty committees and administrators to make decisions about merit increases, promotion, and tenure" (Davis, 2009, p. 534). Abdulla, Badri, Dodeen, and Kamali (2006) state that in addition to promotion and tenure decisions, SETs are also used in long-term contracts, contract renewals and award related decisions. Since promotion often requires documentation of effective teaching, SETs are typically used as the sole or at least primary metric (Fike, Doyle, & Connelly, 2010). SET outcomes are also used in faculty hiring, firing and reappointment decisions (Uttl & Smibert, 2017). SETs have become an institutionalized component of the American Higher Education process (Avery, Bryant, Mathios, Kang, & Bell, 2006).

Due to the extensive dependence on SETs and their effect on career advancement, "any potential bias in those student evaluations is a matter of great consequence" (Driscoll, Hunt, & MacNell, 2014, p. 3). Therefore, this review of the literature will include the aforementioned biasing variables, and will explore what we know, what we don't, and what areas are still complex and unclear.

Instructor Variables of Gender and Race

Gender

Gender bias and gender inequality in the workplace are common themes in business and social science literature. While, historically, business has been a male dominated field, gains in staffing equality have reached some business sectors and not others. In 2016, the Peterson Institute for International Economics surveyed 21,980 firms headquartered in 91 different countries. Results indicate 60% (n=13,017) had no female representation on their board of directors, over 50% (n=11,802) had no female "C-suite" executives, and less than 5% (n=985) had a female chief executive officer (Noland, Moran, & Kotschwar, 2016). In top positions in postsecondary business schools, less than 20% of business school deans identify as female (AACSB, 2015), and only 20.2% of professors who hold the rank of full professor identify as female (AACSB, 2016). Specific to academic instructor roles, this disparity has been well documented, where men tend to be regarded as "professors" and women as "teachers" (Miller & Chamberlin, 2000). The role of business professor is still male dominated at a 2-1 male to female ratio. In 2019, AACSB reported that 67% (n=90,554) of all business school faculty identified

as male (Full-time=69%, Part-time=65%)[2].

With this clear male bias in higher education in general and business schools in particular, it would seem to logically follow that female instructors would obtain lower SETs. This is especially troubling since SETs play a significant role in career outcomes for college instructors. However, some researchers, through several different studies, did not find evidence of gender bias. In a summary of related research, Aleamoni, citing 16 studies, "reported no differences between faculty ratings made by male and female students" (Aleamoni, 1999, p. 3). Bennet found that although "students are less tolerant of female instructors whom they perceive as lacking professionalism and objectivity than they are of male instructors who lack the same qualities…the study offers no evidence of direct bias in formal student evaluation of instructors" (Bennet, as cited in Driscoll et al., 2014, p. 294). Basow and Distenfeld (1985) found that teacher expressiveness was a more important factor for males in student evaluation of teaching, regardless of the instructor's gender. In 1992, Feldman concluded that the majority of laboratory and experimental research on college students' preconceptions of male and female instructors showed that students' global evaluations of male and female college teachers as professionals were not different. In 1993, Feldman found that although a majority of studies have found instructor gender is not the cause of differences in student global ratings, when significant differences are found, more of them favor females than males. In 1995, Tatro found that female instructors received significantly higher ratings than male instructors. In 2009, Rush, Shaw and Young noted several researchers (Basow &

[2] Note: Schools had the option to identify personnel gender as Other when applicable. These figures have been excluded from the tables in this guide due to insufficient numbers of such designations to protect data privacy/confidentiality.

Distenfeld, 1985; Feldman, 1983, 1993; Goodwin & Stevens, 1993; Hancock, Shannon, & Trentham, 1993) could also not find evidence of gender differentiations. Basow states that although it is likely that gender is a contributing factor in SETs, the relationship is a complex one (2000).

Conversely, some researchers did find evidence of gender bias. In 2009, McPherson, Jewell and Kim found "Male instructors get better scores than females" (p. 37). And multiple other researchers state evidence claiming SETs are biased against female instructors (Basow, 1994; Basow & Silberg, 1987; Kaschak, 1978; Koblitz, 1990; Martin, 1984; Rutland, 1990; Watchel, 1998). In 1995, Basow found women professors are "expected to be more open and accessible to students as well as to maintain a high degree of professionalism and objectivity. Female instructors who fail to meet these higher expectations are viewed as less effective teachers than men" (p. 4).

There is some indication that when both the gender of the student and instructor are the same, higher outcomes may result on some teaching dimensions (Aleamoni, 1981; Centra, 1981; McKeachie, 1979). Driscoll et al. (2014) summarizing the work of Basow (1995), Centra and Gaubatz (2000), Feldman (1992), and Young, Rush, and Shaw (2009) conclude "students perceive, evaluate, and treat female instructors quite differently than they do male instructors" (p. 2). Specifically, research exists suggesting that the gender of the instructor plays an important role in determining SET scores, and that "students perceive female instructors differently than men" (McPherson, Jewell, & Kim, 2009, p. 6) finding that all other things equal, males received significantly higher ratings than their female contemporaries (McPherson et al., 2009). In fact, when using false instructor gender identities in an online class, "students rated the male identity significantly higher

than the female identity, regardless of the instructor's actual gender, demonstrating gender bias" (Driscoll et al., 2014, p. 1). Additionally, when receiving negative feedback from a professor, students perceived female professors as less competent than male professors (Sinclair & Kunda, 2000).

There is previous evidence that SET responses directly related to teaching effectiveness, are not always a measure of teaching effectiveness (Freishtat & Stark, 2014; Watchel, 1998). In fact, recent research shows "SET are more sensitive to students' gender bias and grade expectations than they are to teaching effectiveness" and "measure students' gender biases better than they measure the instructor's teaching effectiveness" (Boring, Ottoboni, & Stark, 2016, p. 1). Largely, SETs handicap female instructors and "there is no evidence that this is the exception rather than the rule" (Boring, et al., 2016, p. 11). Finally, Boring et al. (2016) warn that until these biases are further investigated and mitigated, "SET should not be used for personnel decisions" (p. 11).

Gaps in the Literature on Gender

As shown above, there is evidence that affirms and rejects the hypothesis that there is gender bias in SETs. As such, any additional research can possibly help resolve these conflicting results. More critically, there is almost no research focusing specifically on faculty in a business school. This is especially valuable since business schools, along with fields like engineering and math, are male dominated. Given the need for clear and consistent evidence, this study was designed to answer the following research question: What is the effect of instructor gender on SET ratings?

Race

Racial bias and racial inequality in the workplace are also common themes in business and social science literature. In 2017, there were only four Black CEOs in the Fortune 500 (White, 2017). In the U.S., more than two-thirds of business professors are Caucasian, 17.5% are Asian/Pacific Islander, but only 4.1% are Black and 2.7% Hispanic (AACSB, 2019).

While there is a wealth of scholarship on the relationship between gender bias and SETs, substantially less has been published on how instructor race and ethnic identity informs SET effectiveness (Anderson & Smith, 2005; Bavishi, Helb, & Madera, 2010; Hawkins & Smith, 2011; Williams, 2007). In the SET literature, there is substantively more research on the effects of gender than race and ethnicity, probably because non-White professors are such a minority in the professorate (Basow, Martin, & Codos, 2013).

Although SETs are designed as objective assessments of teaching effectiveness, extraneous factors such as the instructors' race "affect the composition and educational atmosphere at colleges and universities" (Perry, Wallace, Moore, & Perry-Burney, 2015, p. 29). Depending on the type of course taught, research has found that, based solely on professor race, students' first impressions "do influence their judgments of their professors' perceived bias and, consequently, their perceptions of professors' subjectivity and expertise; thus, a professor's race/ethnicity can directly influence student evaluations" (Littleford, Ong, Tseng, Milliken, & Humy, 2010, p. 242). Holistically, they conclude instructor race does influence student perception of faculty effectiveness, as measured through SETs.

Studies have found that African American and Hispanic faculty obtain lower SET results than Caucasian and Asian faculty (Basow & Martin, 2012). For SETs in general, Caucasian and Asian faculty appear to receive higher evaluations than Black and Hispanic faculty (Hamermesh & Parker, 2005).

Latinx

DiPietro and Faye (2005) studied the SET outcomes of African American, Hispanic, Asian American, and White instructors and concluded that Latinx faculty received the lowest ratings. Anderson and Smith (2005) found "ratings of professor warmth and availability for Latino professors appear to be contingent on their teaching style, whereas the rating of Anglo professors' warmth is less contingent upon teaching style" (p. 196). Additionally, "Latina professors were viewed as more warm when they had a lenient teaching style and less warm when they had a strict teaching style when compared with Anglo women professors with respective styles" (p. 184). Lastly, even Latinx students rated Caucasian professors as more capable than Latino professors (Anderson & Smith, 2005).

African American / Black

In the professoriate, African American often have higher service requirements than their White peers (Griffin & Reddick, 2011). African American professors are rated as less legitimate and competent than Caucasian and Asian American professors (Bavishi et al., 2010). Harlow (2003) went as far as to suggest that black professors' work in the classroom is dissimilar and includes added complexity over that of their Caucasian colleagues. Navigating a career with a devalued racial status requires widespread emotional management. Harlow also suggests these differing job performance

expectations influenced by race can also "affect the negotiation of self and identity in the classroom, influencing the emotional demands of teaching and increasing the amount of work required to be effective" in the undergraduate college classroom (p. 348).

In a 2010 study, Bavishi et al. found African American professors were viewed by college preparatory students – before they even met – as less competent and legitimate compared to Caucasian professors. Indeed, lower status groups must "prove" their competence when evaluated for professional positions (Foschi, 2000).

However, other research on SET racial bias has also proven inconclusive. In 2005, Anderson and Smith found no student same-ethnicity preference, meaning students did not, as measured per SET outcomes, rate same-ethnicity professors higher. Williams' research concluded "students are apt to give low evaluations of teaching effectiveness" minority instructors when the course content is controversial (2007, p. 170). Littleford et al.'s (2010) findings suggest that, when teaching a race-focused diversity course, African American instructors might have an advantage over their European American colleagues. Interestingly, a 2011 study of tenure-track faculty at a Predominantly White Institution (PWI) found that undergraduates rated faculty identifying as "Other" higher in "overall value of the course" and "overall teaching ability" than faculty identifying as Caucasian on mean SET outcomes; both Caucasian and "Other" were substantially higher than African Americans (Hawkins & Smith, 2011).

A study by Basow et al. (2013), the first of its kind, examined the effects of professor race after hearing a computer-based lecture given by both a Caucasian and African American computer-animation professors. In this simulated experience, African American professors were rated higher than Caucasian professors on their hypothetical

interactions with students. Assessment results, however, showed students scored better – i.e., learned more – from a Caucasian professor-delivered lecture than they did from an African American professor-delivered lecture. Assessment results may be due to students paying more attention to the more normative professor. "If a White male professor is viewed as the most normative and credible, especially for a technical lecture, students may pay more attention and therefore score higher on the recall test" (Basow et al., 2013 p. 355).

Gaps in the Literature on Race

Hawkins and Smith (2011) stated "in the absence of empirical research, Black faculty's assertions are neither supported nor refuted; rather they are simply anecdotal" (p. 150). The above stated results, both founded and unfounded, proves that racial bias evidenced through SETs is situational, needs to be delineated, and generalizations should be made with caution. Littleford et al.'s (2010) findings "highlight the importance of measuring multiple domains rather than one global indicator when having students evaluate instructors of diversity courses" (p. 242); their outcomes highlight the need to keep measuring factors separately, as this "will help researchers and educators better understand the multiple domains of students' evaluation" (p. 242). However, when factors within an African American/black faculty member's control have been addressed yet negative SET outcomes persist, "then a part of the students' evaluation of teaching may rest in the students' biases, prejudices, misperceptions, and lack of interaction and experience with such faculty" (Stewart & Phelps, 2000 as cited in Hawkins & Smith, 2011, p. 159).

Although the few studies that have been conducted have found that Hispanic and Asian American faculty were rated lower than Caucasian faculty (Anderson & Smith, 2005; DiPietro 2005; Hamermesh & Parker, 2005), only a paucity of research has focused specifically on African American faculty (Hawkins & Smith, 2011). More research on SET bias for African American faculty is needed, but in the meantime "race should not be ignored as a factor when student evaluations are used as a criterion for administrative decisions regarding merit, promotion, and tenure" (Hawkins & Smith, 2011, p. 159).

Also, Latino students rated Caucasian female faculty as more capable than Caucasian male faculty (Anderson & Smith, 2005). "Because of the dearth of studies on ethnic bias with Latinos as targets or participants, future work may help to explain these particular patterns" (Anderson & Smith, 2005, p. 199).

As shown above, there is evidence both for and against racial bias in SETs. As such, any additional research can possibly help resolve these conflicting results. More critically, there is almost no research focusing specifically on African American and Hispanic faculty in a business school. This is especially valuable since business schools, along with fields like engineering and math, are Caucasian dominated. What is the effect of this on SET ratings? This is another of the questions that this dissertation was designed to answer.

Course Factors of Class Size and Course Content

Class size

In the context of the United States, existing SET research has only focused on narrow environments (one school or one subject), frequently does not address bias, and has also found conflicting results (Bettinger & Long, 2018).

College class sizes and the student-to-faculty ratio have been steadily increasing for over five decades (Latta & Warren, 1978). In college-level courses, does size matter? Parents believe so, as they are willing to shell out much larger tuition premiums to ensure smaller class sizes; and students dislike the larger class sizes (Bedard & Kuhn, 2008). A 1979 study concluded large class sizes had a negative effect on student performance (Glass & Smith). Large class sizes can decrease student-faculty communication, both in and out of the classroom, and affect student-faculty relationships (Bettinger & Long, 2018; Ting, 2000). It can affect something as simple as not knowing the students' names or something as substantial as not having enough bandwidth to give personalized, itemized feedback for qualitative assignments. In fact, timely feedback on class assignments and instructor-student interactions were both significantly correlated with perceptions of course effectiveness in an Introduction to Managerial Accounting course (Elson, Gupta, & Krispin, 2018). Using a representative sample that encompassed economics courses at all undergraduate college levels, Bedard and Kuhn (2008) found a consistently large, non-linear, and statistically significant negative impact of class size on SET outcomes. These findings were replicated in a study conducted by Mandel and Sussmuth (2008).

Differentiated or personalized instruction in large classes becomes much more difficult, and as personalized instruction decreases, the students' sense of belonging in the class, interest in the material, and the motivation to master it, can wane (Scheck, Kinicke, & Webster, 1994). Active participation may also suffer, since discussion as a teaching method is not practicable (Ting, 2000). Student participation, which contributes to a positive teaching and learning atmosphere, is limited in large class sizes (Durden & Ellis, 1995). McPherson (2006) found that class size affects SET outcomes: teaching smaller size classes resulted in better SET scores, ceteris paribus, possibly because of the "commonly held view that teaching is most effective in relatively small class sizes" (p. 18). McCahan and Rolheiser (2018) found larger course sizes were also associated with lower Institutional Composite Mean (ICM) on SET outcomes. Individually and collectively, these factors contribute to SET outcomes.

Conversely, other studies did not find direct evidence of SET outcomes being affected by large class sizes. An earlier study from 1977 determined there were no significant differences in SET outcomes from students due to size alone in small, medium, and large classes (Aleamoni & Graham). In 2006, McPherson concluded the "findings of researchers in this area are varied and sometimes in opposition to each other" (p. 4). Ting (2000) found that while class size did reduce students' self-rated studying efforts, it did not have a negative effect on student satisfaction with the course. McPherson et al. (2009) found that while class size and instructor experience are important factors in determining SET outcomes from lower division classes, this effect was not found in upper-level courses. McPherson (2006) also found evidence of possible

contamination of SET outcomes due to instructors influencing the scores by raising grade expectations.

Gaps in the Literature on Class size

This literature has found larger class sizes decrease SET outcomes, which could be a reflection of how learning works in large versus small class sizes. Additional studies found no effect on SET due to large class sizes. This study aimed to find whether SET has a positive effect, which is not supported by existing literature, or if class size has no effect or a negative effect on SET outcomes.

Qualitative to Quantitative Ratio

In a summation of studies related to faculty gender bias, Arreola (2006) insinuates that the bias may be due to the types of courses instructors teach. There are many variables that can affect students' perception of course content, design, and delivery, and overall SET outcomes. These include time of day, class size, delivery mode, workload, required versus elective, course level, and the quantitative versus qualitative nature of the course.

Ramsden (1991) found the course subject matter area does have a significant effect on SET outcomes. Watchel (1998) found that teachers of non-required elective courses have higher SET ratings than teachers of required core courses. Aleamoni (1999), citing nine studies on the effect of elective vs required courses, also concluded non-required elective courses have higher SET ratings. SET outcomes are typically higher for advanced courses than beginner courses, and typically higher for electives than required courses (Cranton & Smith, 1986; Feldman, 1978; Marsh, 1980, 1983; Murray et al., 1990, Ting, 2000). Ting (2000) found that "courses dealing with abstract ideas or

numbers are less popular than others with more concrete contents; it typically requires better skills and careful planning to make these subjects both attractive and interesting to students" (p. 643).

Studies also found that course level impacts SET outcomes, mentioning that student age and maturity level could impact outcomes (Aleamoni, 1999; Feldman, 1978; Marsh, 1987; Watchel, 1998). Additionally, Uttl and Smibert (2017) concluded the course subject is strongly associated with SET results and has a considerable influence on how instructors are evaluated by administration. In a study involving 14,872 SET ratings, Uttl and Smibert (2017) found instructors teaching quantitative courses are less likely to "receive tenure, promotion, and/or merit pay when their performance is evaluated against common standards" (p. 1). Still, other studies conclude that the discipline of the course had no effect on SET outcomes (Aleamoni, 1999; Cashin, 1990).

Additional research spanning 18 years suggests quantitatively rigorous courses rank much lower on SETs than other non-math or science courses (Watchel, 1998). At the section-level, Uttl and Smibert (2017) found English course SETs were much higher than math course SETs, with 71 percent of English courses, and just 21 percent of math courses, passing the overall mean as the standard. Substantial evidence exists indicating SETs in both business and non-business courses are affected by the course's subject matter, with prior research suggesting course content might explain variation in SET results across disciplines (Abdulla et al., 2006). In an analysis of SET results by business disciplines, Marsh and Overall (1981) found the course type – quantitative versus qualitative – to be an insignificant variable. However, several studies show arts and humanities courses in general have higher SETs than business courses (Basow &

Montgomery, 2005; Cashin, 1990; Ory, 2001). Additional research spanning decades has identified positive correlations between interest level and SET ratings (Bonitz, 2011). In reference to response rates and self-selection of responses, Kherfi's (2011) findings confirm students who do better in a course are more likely to complete the SET, although this would divest itself across business disciplines and therefore does not speak to the qualitative versus quantitative nature of the courses in this study. Lastly, Uttl and Smibert (2017) found that instructors who teach quantitative courses are less likely to receive teaching awards.

Gaps in the Literature on Course Content: Qualitative to Quantitative Ratio

So, the question is not whether course content affects SET outcomes; the question is how and in what situated contexts? Of particular relevance to this study, almost all majors offered by business schools require both quantitative and non-quantitative courses. As such, a business school is an ideal setting to test whether this variable has an effect on SET rating. This study examined whether instructors teaching quantitative courses are negatively affected in their SET outcomes, based solely on the quantitative nature of the business course(s) they teach.

Student Factor of Section-Level Grade Point Average

GPA

It has long been assumed that giving higher grades leads to better student evaluations. It has been suggested that high course satisfaction can cause high grades and that leniency is irrelevant (Howard & Maxwell, 1982). However, a recent study by Stroebe (2017) suggests grading leniency is caused by SETs, suggesting a GPAs rise in the 1980s coincides with the use of SETs as a measure of faculty evaluation became

standard practice. There is no clear evidence that students augmented the time they spend on their academics, but there is clear evidence that over time students are spending less time on their academics.

There is evidence to indicate "that instructors can buy higher SET scores by awarding higher grades" (McPherson, 2006, p. 3). Gorry's (2017) results suggest increasing GPA increases SET outcomes in a sliding fashion, meaning the higher the GPA lift the higher the SET outcomes. This was implemented by a grade predetermined ceiling.

Incentives to Increase Response Rate of Online SETs

While data regarding whether, and exactly how, instructors are influencing SET outcomes are not included in this study, this issue is common in the literature. Incentives are used to influence behavior across the business and educational landscapes. For purposes of this literature review, incentives are described as the mechanisms and manner used to influence student behavior to complete the SET, and also complete SETs favorable to the instructor. However, incentives also include the mechanisms and manner institutional administration uses to influence student response rate of SETs.

Reisenwitz (2016) found that incentives, such as offering early registration times, withholding access to grades, and increasing grades, improve response rates. Although Porter (2004) asserts incentives conditional on survey completion have no effect on response rates, other research states otherwise. Cobanoglu and Cobanoglu (2003) suggest using incentives for completion, such as prizes or extra credit for the course. Online response rates were generally lower than the in-class response rates, unless a grade

incentive was used to increase response rates (Dommeyer, Baum, Hanna, & Chapman, 2004)

Having faculty instill in students that their feedback is valuable and put to good use to make course improvements is paramount. Open-ended questions provide actionable feedback, where Likert scale questions do not. To increase student participation in the evaluations of faculty members, university administrators must learn students' attitudes toward current and proposed methods for gathering teaching evaluations (Dommeyer et al., 2004).

In 2012, the institution in this study added the lure of seeing SET outcomes for professors in upcoming semesters, allowing students to view aggregate results of the following four SET questions:

- The instructor (or lab/recitation instructor) provided useful feedback about exams, projects, and assignments (labeled as "Feedback").
- So far, the instructor (or lab/recitation instructor) has applied grading policies fairly (labeled as "Grading").
- The instructor (or lab/recitation instructor) taught this course/lab/recitation well (labeled as "Teaching").
- I learned a great deal from this course/lab/recitation (labeled as "Learning").

The rationale is that publicizing results to students, even in a limited fashion, will increase response rates and provide students with assistance when selecting courses for the upcoming semester. Scriven (1981) argued "that it is unethical to deny students the opportunity to view the results of ratings which they have engendered" (as cited by Watchel, 1998, p. 205). In fact, back in 1993 at Northern Illinois University a student

association organized a boycott of SETs due to administration's refusal to share the outcomes. On the contrary, faculty have voiced opinions that SET data should not be publicly available (Chan et al., 2014).

Gaps in the Literature Students' Section-Level Grade Point Average

Griffin, Hilton, Plummer, and Barret (2014) found only a moderate correlation between SETs and GPAs in 2073 general education religion courses at a large private university. They also found some negative correlations when parsing into individual courses and instructors. This study examined whether there is a significant correlation between section-level GPAs and SETs, but in the setting of in a large, public business school.

Data Collection Factor of Paper-and-Pencil vs Online

Although the electronic administration of surveys is a recent technique, electronic surveys have been compared with traditional survey administration methods and been proven effective across a variety of organizational applications such as market research, personnel evaluations, and psychological counseling (Rosenfeld, Booth-Kewley, & Edwards, 1993; Sproull & Kiesler, 1991). Researching SETs, in various forms, began in the 1920's (Brandenburg & Remmers, 1927), although electronic SET surveys have only been used since the early 1990s. Conducting SETs through an online process has many benefits for the budget, the environment, and can be more efficient and effective (Fike et al., 2010).

Researchers found no significant differences in SET outcomes between the traditional pencil and paper and online evaluation formats (Donovan, Mader, & Shinsky, 2006). Studies also found even when response rates lagged for online SETs results were

similar for both methods (Avery et al., 2006; Fike et al., 2010).

Research has also shown students clearly prefer the online model (Layne, DeCristoforo, & McGinty, 1999; Donovan, Mader & Shinsky, 2007), and students' enthusiasm to complete a SET survey is directly related to their satisfaction with the SET process (Abbott, Wulff, Nyquist, Ropp, & Hess, 1990; Dommeyer, Baum, & Hanna, 2002). When completing online SETs, students wrote more comments, comments were longer in length, and qualitative detail was more often formative than summative (Donovan et al., 2006). Formative comments are more likely to offer actionable instructor feedback as compared to summative comments.

There are four primary variables in the SET literature for SET data collection and modality: response rates, incentives, selection bias, and anonymity.

Response Rates

The primary empirical discussion around switching from pencil and paper to online was based on how the data collection modality would/could change the response rates.

Response rates are often contingent upon the faculty and administration's role in the student "buy in" process (Fike et al., 2010). Sax, Gilmartin and Bryant (2003) suggest completing too many surveys can cause students to hit a saturation point and stop completing surveys altogether. Typically, response rates have dropped during the migration to online data collection, and the most common ways to increase response rates is to repeatedly send reminders (Crawford, Couper, & Lamias, 2001).

Avery et al. (2006) found, despite the lower response rate for online SETs, average online SET outcomes were not different between online and pencil and paper

SET outcomes. However, Mau and Opengart (2012) concluded faculty do receive higher SET outcomes using an in-class or paper-based assessment.

Nested inside of the response rate factor is the concept of response bias. Response bias and nonresponse bias arise when there are differences in significant variables between respondents and non-respondents. Two studies have found significant differences between undergraduate business students who completed the online SET versus those who did not complete the online SET (Estelami, 2015; Reisenwitz, 2016). Reisenwitz (2016) found "persistent significant differences regarding gender, race, and GPA between students who participate in online student evaluations and those who do not" (p. 14).

Selection Bias

Many studies have found that respondents differ from non-respondents in their behaviors and attitudes (Goyder, 2001). Layne et al. (1999) found that students with higher GPAs were more likely to complete SETs, and second-year students were more likely than seniors to complete SETs. Avery et al. (2006) found that students in smaller sections were also more likely to complete SETs.

Fike et al. (2010) found that in an online data collection SET system, administration needs to communicate to instructors that their efforts of encouragement to students regarding SET completion create the student buy-in needed for effectiveness. Estelami (2015) concluded that high performing students were more likely to complete SETs. Although response rates are included in this study, student data were not included and therefore selection bias was not tested.

Anonymity

The issue of anonymity was always an issue with SETs. Whether the instructor is in the class or not can also affect SET results, with some findings indicating higher scores when the instructor is present (Feldman, 1979). Although over 30 years ago, Pulich (1984) suggested eliminating student anonymity from the SET process, her conclusions were not based on empirical evidence and are commonly disputed as Watchel (1998) states "most authors (for example, Braskamp et al. 1984; Centra, 1993; McCallum, 1984) recommend that student raters remain anonymous" (p. 195). "Feldman (1979) and Blunt (1991) report that students tend to give somewhat higher ratings when they identify themselves compared to those when they remain anonymous" (Wachtel, 1998, p. 195).

Prior to digital submission of student deliverables, students worried their handwriting would be recognized; Layne et al. (1999) found some students preferred the online method because of the lack of anonymity using their handwriting. Additionally, the online method is preferable to students because of the convenience and lack of time limits; this allows for more formative, constructive responses to open-ended survey questions (Fike et al., 2010).

Assuring students that their responses are anonymous dismisses fears and creates SET buy-in (Fike, et al., 2010). Dommeter et al. (2004) found no significant difference between paper and pencil and online mean SET outcomes, despite offering incentives to the online SET respondents.

The students included in this study submitted their responses anonymously, for both paper-and-pencil and online data collection methods.

Summary

While there is an abundance of literature on most SET variables, much of it is often contradictory. Additionally, much of the research has used single variables rather than several variables together. Minimal past research has examined bias in SET ratings in the context of an academic school or college within a larger university setting. Therefore, this study addressed this gap in the literature by examining bias in SET ratings within the context of a school of business in a large, urban, research university. Given its scope and aim, this study may serve as the first of its kind and will add to the literature on SETs.

CHAPTER 3

METHODOLOGY

Study Setting and Research Design

This study was conducted at a large, urban, public university located within the Atlantic Corridor of the United States. The dataset was compiled by merging SET outcomes, section-level GPA results, and faculty demographic data used for accreditation purposes. A six-year longitudinal dataset was assembled using internal business school data including Instructor demographics, section-level course demographics, section-level student GPAs, and section-level SET/eSET outcomes.

- Instructor demographics: At the point of hire, faculty complete a demographic information form that includes race and, among other variables, their self-identified gender: choices are limited to male and female.
- Course demographics: This variable includes, among other variables, faculty subject matter experts' qualitatively coded labeling of the course as having one of the following three skillset functions: (1) primarily qualitative, (2) blend of qualitative and quantitative, and (3) primarily qualitative.
- Student GPAs: Section-level GPAs were added to the dataset for each corresponding course that valid SET/eSET data were obtainable. Students who earned an Incomplete (I) or Withdraw (W), and courses labeled pass/fail courses were not included in the study.
- SET/eSET data: A six-year dataset was compiled and consisted of all fall and spring courses, between fall 2009 and spring 2015, that met the standards for inclusion.

Sample

Summer courses were not included in the study because instructors who teach summer courses are often limited to non-tenure-track faculty members and PhD students teaching for the first time, courses are delivered in a compressed delivery format consisting of only six weeks instead of the traditional 14-week semester, and often include section sizes under the threshold allowable in fall and spring semesters (i.e., $n<10$). Moreover, students are often atypical by way of taking courses for a second or third time because of failure(s), Withdraw(s) or Incomplete(s), or being a new transfer to the institution. Due to these differing demographics, the SET/eSET outcomes were deemed atypical and not included in the study. Additionally, graduate-level and independent study class sections were not included, nor were sections where the university threshold of at least eight student SET/eSET responses per section were recorded. Although data were available on both undergraduate and graduate students, it was decided to include only data from undergraduates. This is consistent with a majority of the published research and reflects that fact that graduate students are typically older and more mature; as such it is assumed they interpret the SET questions differently.

Instrumentation

The dataset for this study is comprised of SFF data spanning fall 2009 through spring 2015; however, half way through this six-year period SFF data collection went from in-class, paper-and-pencil model, to the current online data collection model. As a result, the first three research questions of instructor gender and race, course size and qualitative/quantitative course nature, and section-level GPA, will span between the effect of in-class, paper-and-pencil and online models; analyses will be run together and

by splitting the two, three-year time periods of fall 2009 to spring 2012 (paper) and fall 2012 to spring 2015 (online). The SET form includes the following questions:

Table 3.1:

Complete List of SET Questions

SET Questions NOT Used in This Study	Rating
Before enrolling, my level of interest in the subject matter of this course was:	Low, Moderate or High
Expected grade is:	A, B, C, D, F
Course is:	Required or Elective
On average, hours per week spent preparing for class and completing course assignments:	<1, 1-2, 2-3, 3-4, 4-6, 8+
I came well prepared for class.	See Likert Scale below
What aspects of the course or the instructor's approach contributed most to your learning?	Descriptive text response.
What aspects of the course or the instructor's approach would you change to improve the learning that takes place in the course?	
Please comment on the instructor's sensitivity to the diversity (for example, political viewpoint, race, ethnicity, national origin, gender, sexual identity and disability) of the students in the class	
SET Questions USED in This Study	Rating
Q1: The instructor clearly explained the educational objectives of this course.	Rate each item below using the scale to the left where SA = Strongly Agree and SD = Strongly Disagree. If an item does not apply, select N/A (Not Applicable).
Q2: The instructor was well organized and prepared for class.	
Q3: The instructor was conscientious in meeting class and office hour responsibilities.	
Q4: The instructor promoted a classroom atmosphere in which I felt free to ask questions.	
Q5: The instructor provided useful feedback about exams, projects, and assignments.	
Q6: So far, the instructor has applied grading policies fairly.	
Q7: The instructor taught this course well.	
Q8: The course content was consistent with the educational objectives of this course.	
Q9: The course increased my ability to analyze and critically evaluate ideas, arguments, and points of view.	
Q10: I learned a great deal in this course.	

There are also three open-ended questions on the SET form which were also excluded. For all research questions and variables, the SET questions were ran separately and collectively, collectively using a composite mean. Because the 10 evaluative questions us he same five-point Likert scale where Strongly Agree=5, Agree=4, Neutral =3, Disagree=2, and Strongly Disagree=1, a composite mean is possible by adding the 10 questions together and dividing by 10. The composite mean is validated by the results of an Exploratory Factor Analysis using Varimax rotation, which loaded all 10 questions into a single factor (see Chapter 4).

Research Question #1: Gender and Race

Gender

When instructors begin employment at the business school they complete a demographic information form that includes race and, among other variables of citizenship, country of origin, and degrees completed, their self-identified gender (choices are limited to male and female). For the purpose of this study, gender is a binary item where, at the section-level analysis, 67.5% of instructors are male and 32.5% are female. Hancock et al. (1993) found differences between SET outcomes of male and female instructors occur on the individual questions.

Therefore, a two-way repeated measures ANOVA (gender) was conducted to find significant differences between males and females on the individual SET questions.

However, because the data in this proposed study are only obtainable at the section-level, students' gender identification could not be accounted for in this study. Therefore, the gender bias is focused on the instructor's gender, which was gathered from the faculty personnel record form. Completed at the point of hire, faculty are asked in a

binary fashion whether their gender, referred to as "sex," is either male or female.

Race

There are six options for faculty race: African American (not Hispanic origin), American Indian, Alaskan Native, Asian, Pacific Islanders, Hispanic, White (not Hispanic origin), and Other. The section-level analysis revealed 78% of all sections are taught by White instructors, followed by 13.8% by Asian, 5.7% African American, 1.3% American Indian/Alaskan Native, 0.8% Hispanic, and 0.4% Other. Because of the small sample sizes for American Indian/Alaskan Native, Hispanic, and "Other" instructors, this study only focused on White, Asian, and African American demographics. American Indian/Alaskan Native, Hispanic and Other were recoded into a new variable called Other which accounts for 2.5% of the sample.

A two-way repeated measures ANOVA (race) was conducted to find significant differences as a function of race on the individual SET questions. A two-way ANOVA tests the two main effects (gender or race by question) and then the interaction between those two factors.

It should be noted that while there are 5,701 data points in the dataset, there are not 5,701 different instructors. In fact, there are several instructors that are in the dataset over 40 times because they taught multiple courses and course sections between 2009 and 2015. One way to handle this issue is to use the instructor as the unit of analysis, rather than the course section. Both sets of analyses will be conducted for this research.

Research Question #2: Class Size and Content

Class size

Class sizes at the business school range from one to over 500 students in a single course. Larger sections are typically required courses. For the purpose of this study, instead of using the "after two weeks drop/add measure," class size was determined by counting final A through F letter grades given, with Withdraws and Incompletes excluded. To test whether class size has a relationship to SET outcomes, bivariate correlational analyses were conducted between each of the ten questions individually, as well as the composite mean. Since some previous research has found that this relationship is non-linear, both quadratic and cubic analyses were conducted in addition to the typical linear analysis.

Content: Qualitative, Quantitative, or Both

Instructors reported the code for the quantitative/qualitative nature of the course where 1=Qualitative, 2=Blend of Qualitative and Quantitative, and 3=Quantitative. In this study, the variable uses faculty subject matter experts' qualitatively coded labeling of the course as having one of the following three skillset functions: (1) primarily qualitative, (2) blend of qualitative and quantitative, and (3) primarily qualitative. The issue of course content is broadly defined and usually based on assumptions. Courses are typically designated quantitative or qualitative, without "a little of both" option, and this was presumed by department or by program. For example, a course on the ethics of accounting would be labeled quantitative because it is housed in the accounting department, but the course could be taught by a lawyer, and use written communication. Another example would be Management Information systems (MIS) courses which

topics can range from programming to IT architecture. For this study, the subject matter experts' qualitatively coded labeling of the course was done through rigorous discussions between the researcher and the instructors teaching the courses. Each course was held to the same standards for coding due to the calibration of the researcher.

Two answer the content portion of the question, a two-way repeated measures ANOVA (content by question) was conducted to find significant differences as a function of content on the individual SET questions.

Research Question #3: GPA

For purposes of this study, grade point average (GPA) is the section-level mean of all letter grades given where A=4, A-=3.67, B+=3.33, B=3, B-=2.67, C+2.33, C=2, C-=1.67, D+=1.33, D=1, D-=.67, F=0. To test whether section-level GPA has a relationship to SET outcomes, bivariate correlations were run between each of the ten questions individually, as well as the composite mean.

Research Question #4: Data Collection Factor of Paper-and-Pencil vs Online

The dataset for this study is comprised of SFF data spanning fall 2009 through spring 2015; however, half-way through this six-year period SFF data collection went from in-class, paper-and-pencil model, to the current online data collection model. The possible effect of this change was tested through the same two-way, repeated measures ANOVA described above.

CHAPTER 4

RESULTS

The purpose of this study was to investigate whether the instructor race and gender, course content and size, student section-level GPA, and the data collection method – shown to be biasing in the literature – are also significant factors in a business school housed within a large, urban, public, R1[3] institution. Using the SET questions as the dependent variable(s), there were four research questions. Each of the four research questions will be answered and a summary will be presented at the end.

In the analyses of SET data, one of the reoccurring issues concerns the unit of analysis. In general, this is handled two ways: for any analysis that involves instructor characteristics (e.g., gender or race) the appropriate unit of analysis is the individual instructor. For any analysis that involves the course, the unit of analysis is the individual course section. Both of these types of analyses were employed in this dissertation. For instructor analyses (specifically Research Question # 1), the SPSS function termed "Aggregate" was employed. This sums all of the data for each instructor and computes the mean. For these analyses the total sample size is 547. For the remaining research questions, the data for each course section were used as the unit of analysis. For these analyses, the sample size is 5,701.

Descriptive Data on the Sample

Descriptive Statistics for Gender and Race are included in the following Tables.

[3] R1=Research 1 Doctoral Universities (Very high research activity). This is the highest Carnegie research activity classification.

Table 4.1

Descriptive Data for Instructor Gender, Frequencies and Percentages

Gender	Frequency	Percent
Female	182	33.3
Male	365	66.7
Total	547	100.0

Table 4.2

Descriptive Data for Instructor Race, Frequencies and Percentages

Original Race		
	Frequency	Percent
African American (not Hispanic origin)	30	5.5
American Indian/Alaskan Native	2	0.4
Asian/Pacific Islanders	119	21.8
Hispanic	11	2
Other	3	0.5
White (not Hispanic origin)	382	69.8

As shown in Tables 4.1 and 4.2, approximately two-thirds of the sample is male with a slightly higher parentage (69.8%) being White. Since some of the racial groups are too small for analyses (specifically, American Indian/Alaskan Native, and Hispanic, accounting for 2.5% of the sample) they were grouped together and labeled as "Other." The recoded variable that will be used for all subsequent racial group analyses is presented in Table 4.3.

Table 4.3

Descriptive Data for Recoded Instructor Race, Frequencies and Percentages

Recoded Race	Frequency	Percent
African American	30	5.5
Asian	119	21.8
White	382	69.8
Other	16	2.9
Total	547	100.0

Although there are 11 questions in the SET form used by the University, Question #1 – I came well prepared for class – is excluded by the business school used in this research. This question was excluded because it is not a measure of teaching. As such, the SET scale contains 10 questions in a five-point Likert format. Correlations among these questions are contained in Table 4. 4.

Table 4.4

Correlation Matrix for SET Questions

		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Q1	Pearson Correlation	1									
	Sig. (2-tailed)										
	N	5701									
Q2	Pearson Correlation	.894**	1								
	Sig. (2-tailed)	0.000									
	N	5701	5701								
Q3	Pearson Correlation	.811**	.827**	1							
	Sig. (2-tailed)	0.000	0.000								
	N	5701	5701	5701							
Q4	Pearson Correlation	.808**	.740**	.768**	1						
	Sig. (2-tailed)	0.000	0.000	0.000							
	N	5701	5701	5701	5701						
Q5	Pearson Correlation	.876**	.824**	.803**	.824**	1					
	Sig. (2-tailed)	0.000	0.000	0.000	0.000						
	N	5701	5701	5701	5701	5701					
Q6	Pearson Correlation	.799**	.736**	.738**	.755**	.853**	1				
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000					
	N	5701	5701	5701	5701	5701	5701				
Q7	Pearson Correlation	.933**	.882**	.807**	.845**	.891**	.809**	1			
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000				
	N	5701	5701	5701	5701	5701	5701	5701			
Q8	Pearson Correlation	.901**	.859**	.801**	.794**	.861**	.817**	.912**	1		
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000			
	N	5701	5701	5701	5701	5701	5701	5701	5701		
Q9	Pearson Correlation	.865**	.806**	.776**	.785**	.835**	.733**	.908**	.881**	1	
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		
	N	5701	5701	5701	5701	5701	5701	5701	5701	5701	
Q10	Pearson Correlation	.883**	.824**	.775**	.789**	.842**	.758**	.922**	.896**	.947**	1
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
	N	5701	5701	5701	5701	5701	5701	5701	5701	5701	5701

**. Correlation is significant at the 0.01 level (2-tailed).

As shown in Table 4.4, the correlations among the questions are universally high with all questions correlating at least .73 (53% overlap) or higher. One possible analysis, therefore, is to sum the 10 questions to create a composite mean. To demonstrate that this is appropriate, a Cronbach's Alpha was computed which was extremely high (α =.979). As an additional analysis, a Principal Components Analysis was computed which produced one factor that accounted for 84.8 percent of the variance. The component matrix is presented in Table. 4.5, with the questions rank ordered in terms of factor loadings.

Table 4.5

Factor Analysis Component Matrix

Component Matrix[a]

		Component
		1
Q7	The instructor taught this course well.	0.968
Q1	The instructor clearly explained the educational objectives of this course.	0.953
Q8	The course content was consistent with the educational objectives of this course.	0.948
Q10	I learned a great deal in this course.	0.939
Q5	The instructor provided useful feedback about exams, projects, and assignments.	0.935
Q9	The course increased my ability to analyze and critically evaluate ideas, arguments, and points of view.	0.928
Q2	The instructor was well organized and prepared for class.	0.912
Q4	The instructor promoted a classroom atmosphere in which I felt free to ask questions.	0.879
Q3	The instructor was conscientious in meeting class and office hour responsibilities.	0.879
Q6	So far, the instructor has applied grading policies fairly.	0.867

Research question #1: Are the Instructor variables of gender and race biasing factors?

To answer the gender portion of the question, two analyses were conducted: the first used the composite mean and the second used the responses from all 10 questions.

The means and standard deviations for the 10 individual questions and the composite are contained in Table 4.6.

Table 4.6

Descriptive Data for Instructor Gender, Means and Standard Deviations

Question	Gender	M	SD	N
Q1: The instructor clearly explained the educational objectives of this course.	Female	4.257	0.396	182
	Male	4.203	0.405	365
	Total	4.221	0.402	547
Q2: The instructor was well organized and prepared for class.	Female	4.329	0.393	182
	Male	4.298	0.409	365
	Total	4.308	0.404	547
Q3: The instructor was conscientious in meeting class and office hour responsibilities.	Female	4.377	0.314	182
	Male	4.335	0.336	365
	Total	4.349	0.329	547
Q4: The instructor promoted a classroom atmosphere in which I felt free to ask questions.	Female	4.411	0.353	182
	Male	4.354	0.389	365
	Total	4.373	0.378	547
Q5: The instructor provided useful feedback about exams, projects, and assignments.	Female	4.170	0.473	182
	Male	4.091	0.448	365
	Total	4.117	0.458	547
Q6: So far, the instructor has applied grading policies fairly.	Female	4.316	0.364	182
	Male	4.278	0.347	365
	Total	4.291	0.353	547
Q7: The instructor taught this course well.	Female	4.168	0.501	182
	Male	4.106	0.507	365
	Total	4.127	0.505	547
Q8: The course content was consistent with the educational objectives of this course.	Female	4.343	0.314	182
	Male	4.295	0.329	365
	Total	4.311	0.325	547
Q9: The course increased my ability to analyze and critically evaluate ideas, arguments, and points of view.	Female	4.145	0.408	182
	Male	4.106	0.415	365
	Total	4.119	0.413	547
Q10: I learned a great deal in this course.	Female	4.142	0.417	182
	Male	4.093	0.424	365
	Total	4.109	0.422	547
Composite Mean	Female	43.568	3.706	1852
	Male	42.222	4.480	3849
	Total	42.659	4.291	5701

The ANOVA results for the composite are presented in Table 4.7.

Table 4.7

ANOVA Results by Gender for Composite Mean

Source	Mean Square	F	Sig.	Partial Eta Squared
Gender	30.358	2.174	0.141	0.004
Error	13.967			

a. R Squared = .004 (Adjusted R Squared = .002)

As shown in Table 4.7, there is no significant difference between males and females on the SET composite score. To conduct the analysis using the individual questions a two-way repeated measures ANOVA was conducted. These results are presented in Table 4.8.

Table 4.8

Results for Instructor Gender, Between and Within

Source	Mean Squares	F	Significance	Partial Eta Squared
		Between Subjects		
Gender	30.358	2.174	0.141	0.004
Error	13.967			
		Within Subjects		
Question	5.369	219.564	0.000	0.287
Question by Gender	0.024	0.981	0.454	0.002
Error	0.024			

As shown in Table 4.8, there is a significant main effect for Question, but no significant interaction. A graph of the data is presented in Figure 4.1. As shown there, while females have higher means on all 10 of the questions, these differences are not enough to produce a statistically significant difference.

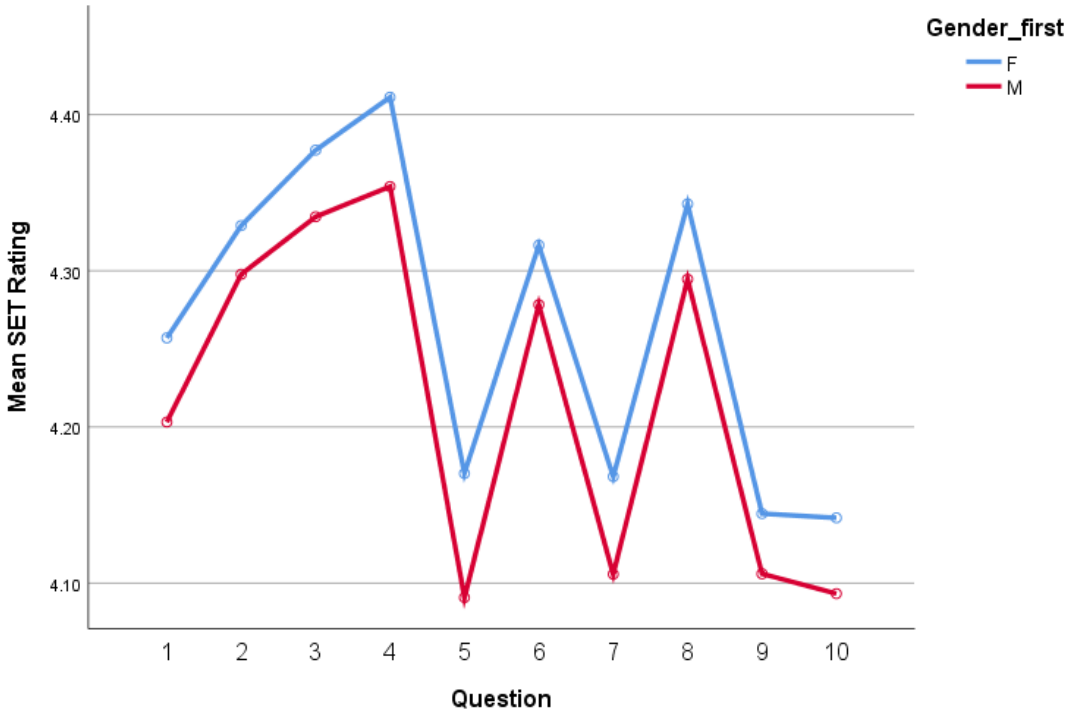

Figure 4.1

Graph for SET Means for Gender

Two identical analyses were conducted to answer the race portion of Research Question #1. The means and standard deviations for the 10 individual questions and the composite are contained in Table 4.9.

Table 4.9

Descriptive Data for Instructor Race, Means and Standard Deviations

Question	Race	M	SD	N
Q1: The instructor clearly explained the educational objectives of this course.	African American	4.084	0.446	30
	Asian	4.044	0.403	119
	White	4.291	0.377	382
	Other	4.121	0.473	14
	Total	4.221	0.403	545
Q2: The instructor was well organized and prepared for class.	African American	4.125	0.442	30
	Asian	4.223	0.391	119
	White	4.350	0.397	382
	Other	4.276	0.437	14
	Total	4.308	0.404	545
Q3: The instructor was conscientious in meeting class and office hour responsibilities.	African American	4.144	0.433	30
	Asian	4.287	0.316	119
	White	4.388	0.311	382
	Other	4.223	0.429	14
	Total	4.349	0.329	545
Q4: The instructor promoted a classroom atmosphere in which I felt free to ask questions.	African American	4.299	0.403	30
	Asian	4.199	0.413	119
	White	4.435	0.346	382
	Other	4.333	0.423	14
	Total	4.373	0.379	545
Q5: The instructor provided useful feedback about exams, projects, and assignments.	African American	3.943	0.521	30
	Asian	3.963	0.467	119
	White	4.186	0.428	382
	Other	3.923	0.614	14
	Total	4.117	0.458	545
Q6: So far, the instructor has applied grading policies fairly.	African American	4.166	0.413	30
	Asian	4.227	0.356	119
	White	4.325	0.338	382
	Other	4.200	0.484	14
	Total	4.292	0.353	545
Q7: The instructor taught this course well.	African American	4.002	0.517	30
	Asian	3.907	0.509	119
	White	4.210	0.477	382
	Other	3.991	0.627	14
	Total	4.127	0.506	545
Q8: The course content was consistent with the educational objectives of this course.	African American	4.211	0.342	30
	Asian	4.213	0.325	119
	White	4.353	0.314	382
	Other	4.218	0.378	14
	Total	4.311	0.325	545

Table 4.9, continued

Q9: The course increased my ability to analyze and critically evaluate ideas, arguments, and points of view.	African American	4.049	0.431	30
	Asian	3.942	0.420	119
	White	4.186	0.382	382
	Other	3.940	0.626	14
	Total	4.119	0.413	545
Q10: I learned a great deal in this course.	African American	4.004	0.448	30
	Asian	3.925	0.424	119
	White	4.182	0.391	382
	Other	3.935	0.624	14
	Total	4.110	0.423	545
Composite Mean	African American	41.0272	4.11982	30
	Asian	40.9308	3.77644	119
	White	42.9057	3.52078	382
	Other	41.2497	4.58603	16
	Total	42.3246	3.74130	547

The results for the composite mean are presented in Table 4.10 and the repeated measures results in Table 4.11.

Table 4.10

ANOVA Results by Race for Composite Mean

Source	Mean Square	F	Sig.	Partial Eta Squared
Race	143.056	10.769	0	0.056
Error	13.284			

Table 4.11

Results for Racial Groups, Between and Within

Source	Mean Squares	F	Significance	Partial Eta Squared
Between Subjects				
Race	143.056	10.769	0.000	0.056
Error	13.284			
Within Subjects				
Question	2.189	93.223	0.000	0.147
Question by Race	0.200	8.533	0.000	0.045
Error	0.023			

As shown in Tables 4.10 and 4.11 there is a significant difference as a function of race which accounts for 5.6 percent of the variance. There is also a significant interaction. The Tukey post-hoc test for race is presented in Table 4.13.

Table 4.13

Tukey Post-Hoc Test on the Main effect for Race

	M	Asian	African American	Other	White
Asian	4.128	-			
African American	4.202	NS	-		
Other	4.222	0.042	NS	-	
White	4.296	0.000	0.000	NS	-

As shown above, White instructors have the highest mean SET scores on the composite, which is significantly higher than Asian and African American instructors. While Asian instructors have the lowest. A graph of the data representing the results for the individual questions is presented in Figure 4.2.

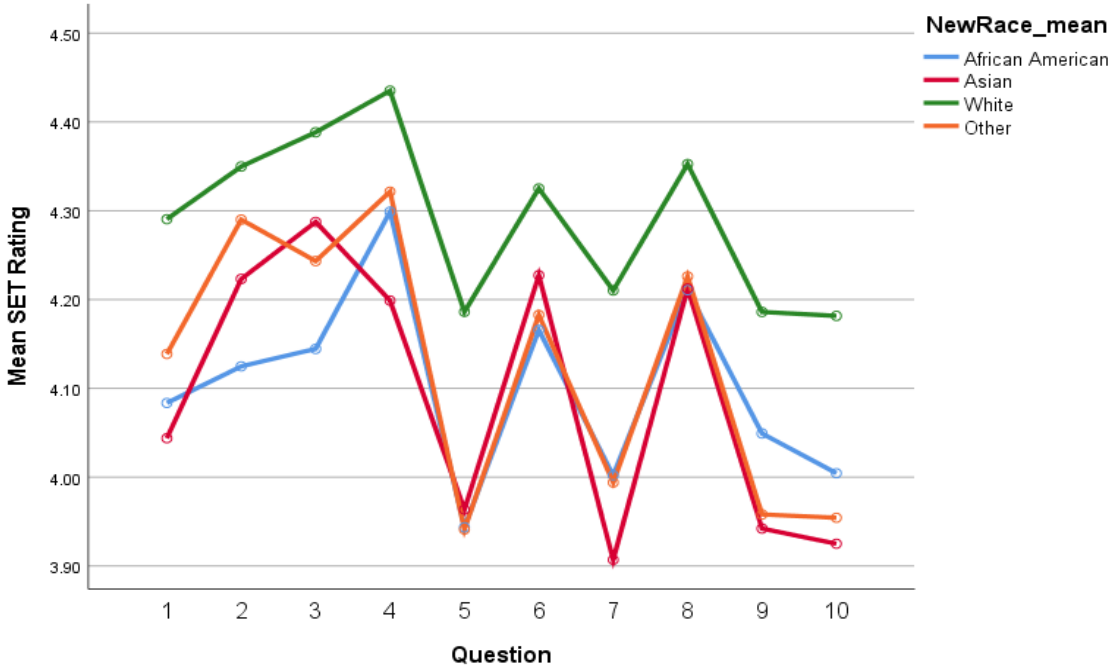

Figure 4.2

Graph for SET Rating by Race

While White instructors have higher means on all 10 SET questions, what is perhaps most interesting about the data displayed in Figure 4.2 is that the difference between the racial groups is not consistent across the ten questions. Questions one (Instructor explained the educational objectives), two (Instructor was prepared for class), and three (conscientious in meeting class/office hour responsibilities) have observable difference among all four racial groups. However, for questions five (Instructor feedback was useful) and eight (Course content was consistent with educational objectives) there was almost no observable variance between African Americans, Asians, and Other. During the time of this study, the business school weighed questions seven (The instructor taught well) and 10 (I learned a great deal) heavily in decision-making. Question seven, the most direct question about Instructor teaching effectiveness, revealed the largest distance in scores among racial groups with White instructors scoring over three tenths higher than Asian instructors. Question 10, the most direct question about student learning, also included a large distance in scores among racial groups with White instructors scoring over two and a half tenths higher than Asian instructors.

Research question #2: Are the Course variables of class size and content (qualitative to quantitative ratio) biasing factors?

To answer the class size portion of the question, bivariate correlations were conducted between (class size) – the number of students in each section – and each of the 10 SET questions. For the purpose of this study, instead of using the "after two weeks drop/add measure," class size was determined by counting final A through F letter grades given, with Withdraws and Incompletes excluded. The correlations between number of students in a section and average SET rating are presented in Table 4.14.

Table 4.14

Bivariate Correlations between SET Question and Class Size

Question	r	sig	r2	Variance Explained	Relationship
Q2: The instructor was well organized and prepared for class.	0.003	0.830	0.000	0.0%	None
Q1: The instructor clearly explained the educational objectives of this course.	-0.013	0.342	0.000	0.0%	None
Q6: So far, the instructor has applied grading policies fairly.	-0.018	0.183	0.000	0.0%	None
Q7: The instructor taught this course well.	-0.021	0.109	0.000	0.0%	None
Q8: The course content was consistent with the educational objectives of this course.	-0.024	0.065	0.001	0.1%	Negligible
Composite Mean	-0.051	0.000	0.003	0.3%	Negligible
Q10: I learned a great deal in this course.	-.052**	0.000	0.003	0.3%	Negligible
Q3: The instructor was conscientious in meeting class and office hour responsibilities.	-.063**	0.000	0.004	0.4%	Negligible
Q5: The instructor provided useful feedback about exams, projects, and assignments.	-.092**	0.000	0.008	0.8%	Negligible
Q9: The course increased my ability to analyze and critically evaluate ideas, arguments, and points of view.	-.092**	0.000	0.008	0.8%	Negligible
Q4: The instructor promoted a classroom atmosphere in which I felt free to ask questions.	-.096**	0.000	0.009	0.9%	Negligible

As shown in Table 4.14, the number of students in a section does not have a meaningful effect on SET outcomes on individual questions or the composite mean. While there are significant correlations, none of these account for more than 1 percent of the variance. As a further analysis, both quadratic and cubic analyses were conducted. There was no clear evidence that either analysis was more accurate than the linear analyses presented in Table 4.14.

To answer the content portion of the question, a two-way repeated measures ANOVA (content by question) was conducted to find significant differences among qualitative and quantitative courses on the individual SET questions. Table 4.15 contains means and Table 4.16 contains results from the repeated measures ANOVA.

Table 4.15

Descriptive Data for Course Content, Means, and Standard Deviations

Descriptive Statistics

Question Number	Content	M	SD	N
Q1: The instructor clearly explained the educational objectives of this course.	Qualitative	4.369	0.416	1719
	Mix of Quant & Qual	4.299	0.434	1594
	Quantitative	4.197	0.473	1750
	Total	4.288	0.448	5063
Q2: The instructor was well organized and prepared for class.	Qualitative	4.426	0.413	1719
	Mix of Quant & Qual	4.395	0.429	1594
	Quantitative	4.308	0.461	1750
	Total	4.375	0.438	5063
Q3: The instructor was conscientious in meeting class and office hour responsibilities.	Qualitative	4.431	0.335	1719
	Mix of Quant & Qual	4.410	0.351	1594
	Quantitative	4.346	0.383	1750
	Total	4.395	0.359	5063
Q4: The instructor promoted a classroom atmosphere in which I felt free to ask questions.	Qualitative	4.507	0.339	1719
	Mix of Quant & Qual	4.367	0.476	1594
	Quantitative	4.277	0.475	1750
	Total	4.383	0.444	5063
Q5: The instructor provided useful feedback about exams, projects, and assignments.	Qualitative	4.221	0.513	1719
	Mix of Quant & Qual	4.141	0.525	1594
	Quantitative	4.104	0.528	1750
	Total	4.155	0.524	5063
Q6: So far, the instructor has applied grading policies fairly.	Qualitative	4.314	0.443	1719
	Mix of Quant & Qual	4.269	0.422	1594
	Quantitative	4.302	0.410	1750
	Total	4.296	0.425	5063
Q7: The instructor taught this course well.	Qualitative	4.329	0.480	1719
	Mix of Quant & Qual	4.219	0.540	1594
	Quantitative	4.066	0.604	1750
	Total	4.203	0.555	5063

Table 4.15 continued

Q8: The course content was consistent with the educational objectives of this course.	Qualitative	4.410	0.354	1719
	Mix of Quant & Qual	4.351	0.370	1594
	Quantitative	4.296	0.383	1750
	Total	4.352	0.372	5063
Q9: The course increased my ability to analyze and critically evaluate ideas, arguments, and points of view.	Qualitative	4.284	0.403	1719
	Mix of Quant & Qual	4.228	0.440	1594
	Quantitative	4.051	0.495	1750
	Total	4.185	0.459	5063
Q10: I learned a great deal in this course.	Qualitative	4.273	0.431	1719
	Mix of Quant & Qual	4.209	0.467	1594
	Quantitative	4.049	0.514	1750
	Total	4.175	0.482	5063
Composite Mean	Qualitative	43.563	3.776	1719
	Mix of Quant & Qual	42.887	4.087	1594
	Quantitative	41.997	4.400	1750
	Total	42.809	4.148	5063

Table 4.16

Results for Course Content, Between and Within Subjects

Source	Mean Squares	F	Significance	Partial Eta Squared
Between Subjects				
Content	106.989	63.718	0.000	0.025
Error	1.679			
Within Subjects				
Question	44.560	1210.774	0.000	0.193
Question by Content	3.388	92.045	0.000	0.035
Error	0.037			

As shown in Tables 4.15 and 4.16, qualitative courses have significantly higher means than quantitative courses on all SET questions and the composite mean. For all questions, the courses labeled as “mixed” have a mean between the quantitative and qualitative courses. The comparison for SET mean scores by content was significant, $F = 92.045\ (2, 5060)$, $p < .000$. Content accounted for 2.5 percent of the variance in SET outcome by course content. The post hoc Tukey test indicated that all three means are

significantly different from each other.

Research question #3: Is the student variable of section-level GPA related to SET ratings?

To test whether section-level GPA has a relationship to SET outcomes, bivariate correlations were run between each of the 10 questions individually, as well as the composite mean. The correlations between the course section-level GPA and average SET rating are presented in Table 4.17 with the correlations rank ordered.

Table 4.17

Bivariate Correlations between SET Question and GPA

Question	r	sig	r2	Variance Explained	Relationship
Q4: The instructor promoted a classroom atmosphere in which I felt free to ask questions.	.343**	0.000	0.118	11.8%	Moderate
Q9: The course increased my ability to analyze and critically evaluate ideas, arguments, and points of view.	.343**	0.000	0.118	11.8%	Moderate
Q6: So far, the instructor has applied grading policies fairly.	.338**	0.000	0.114	11.4%	Moderate
Composite Mean	.333**	0.000	0.111	11.1%	Moderate
Q7: The instructor taught this course well.	.332**	0.000	0.110	11.0%	Moderate
Q5: The instructor provided useful feedback about exams, projects, and assignments.	.328**	0.000	0.108	10.8%	Moderate
Q10: I learned a great deal in this course.	.318**	0.000	0.101	10.1%	Moderate
Q8: The course content was consistent with the educational objectives of this course.	.292**	0.000	0.085	8.5%	Weak
Q1: The instructor clearly explained the educational objectives of this course.	.278**	0.000	0.077	7.7%	Weak
Q3: The instructor was conscientious in meeting class and office hour responsibilities.	.276**	0.000	0.076	7.6%	Weak
Q2: The instructor was well organized and prepared for class.	.206**	0.000	0.042	4.2%	Weak

As shown in Table 4.17, section-level GPA has moderate and weak positive significant relationships between GPA and SET outcomes on individual questions and the composite mean, accounted for up to 11.8 percent of the variance. The composite mean has a moderately significant relationship with section-level GPA, accounting for 11.1 percent of the variance. As an additional way of demonstrating the relationship between section-level GPA and SET ratings, GPA was recoded into six groups (0 – 1.5, 1.56 – 2, etc.). A graph for the composite mean is shown in Figure 4.3.

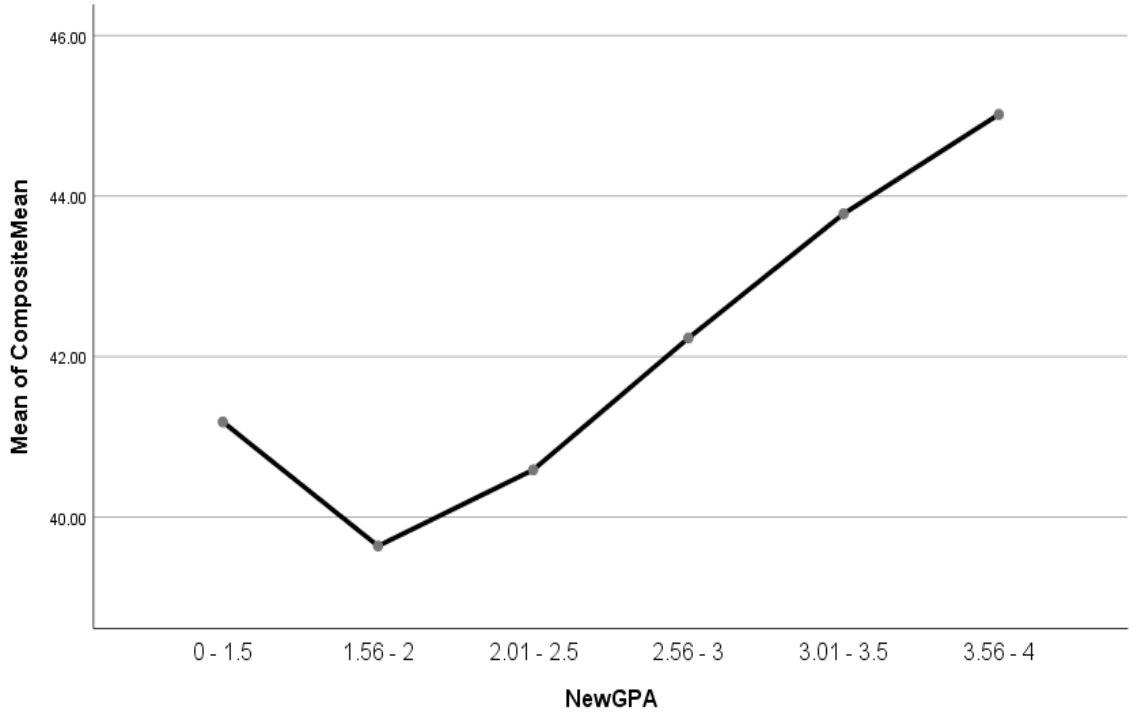

Figure 4.3

SET Ratings by Section GPA for the Composite Mean

Research question #4: Is the data collection method, the change from pencil and paper to online data collection, a biasing factor?

A two-way repeated measures ANOVA (method by question) was conducted to find significant differences among data collection methods on the individual SET questions and composite mean. The results below demonstrate the difference between data collection methods of paper vs online. Table 4.18 contains means and standard deviations and Table 4.19 contains results from the repeat measures ANOVA.

Table 4.18

Descriptive Data for Data Collection Method, Means, and Standard Deviations

Descriptive Statistics

Question	Data Collection Method	M	SD	N
Q1: The instructor clearly explained the educational objectives of this course.	Paper	4.265	0.463	2703
	Online	4.278	0.462	2998
	Total	4.272	0.463	5701
Q2: The instructor was well organized and prepared for class.	Paper	4.362	0.441	2703
	Online	4.348	0.469	2998
	Total	4.354	0.456	5701
Q3: The instructor was conscientious in meeting class and office hour responsibilities.	Paper	4.354	0.368	2703
	Online	4.399	0.384	2998
	Total	4.378	0.377	5701
Q4: The instructor promoted a classroom atmosphere in which I felt free to ask questions.	Paper	4.355	0.458	2703
	Online	4.390	0.459	2998
	Total	4.373	0.459	5701
Q5: The instructor provided useful feedback about exams, projects, and assignments.	Paper	4.137	0.534	2703
	Online	4.140	0.540	2998
	Total	4.139	0.537	5701
Q6: So far, the instructor has applied grading policies fairly.	Paper	4.268	0.440	2703
	Online	4.301	0.430	2998
	Total	4.285	0.435	5701
Q7: The instructor taught this course well.	Paper	4.186	0.575	2703
	Online	4.182	0.571	2998
	Total	4.184	0.573	5701
Q8: The course content was consistent with the educational objectives of this course.	Paper	4.313	0.387	2703
	Online	4.364	0.378	2998
	Total	4.340	0.383	5701

Table 4.18 continued

Q9: The course increased my ability to analyze and critically evaluate ideas, arguments, and points of view.	Paper	4.144	0.485	2703
	Online	4.200	0.457	2998
	Total	4.173	0.471	5701
Q10: I learned a great deal in this course.	Paper	4.147	0.508	2703
	Online	4.174	0.480	2998
	Total	4.161	0.494	5701
Composite Mean	Paper	42.530	4.355	2703
	Online	42.776	4.229	2998
	Total	42.659	4.291	5701

Table 4.19

Results for Data Collection Method, Between and Within Subjects

Source	Mean Squares	F	Significance	Partial Eta Squared
Between Subjects				
Method	8.544	4.644	0.031	0.001
Error	1.840			
Within Subjects				
Question	50.518	1292.811	0.000	0.185
Question by Data Collection Method	0.815	20.855	0.000	0.004
Error	0.039			

As shown in Tables 4.18 and 4.19, the data collection method of paper vs online was significant

$F = 20.855$ (1, 5699), $p < .031$. Statistical significance was mostly due to the large sample size. Course content accounted for one tenths of a percent of the variance. A graph of the results is presented in Figure 4.4.

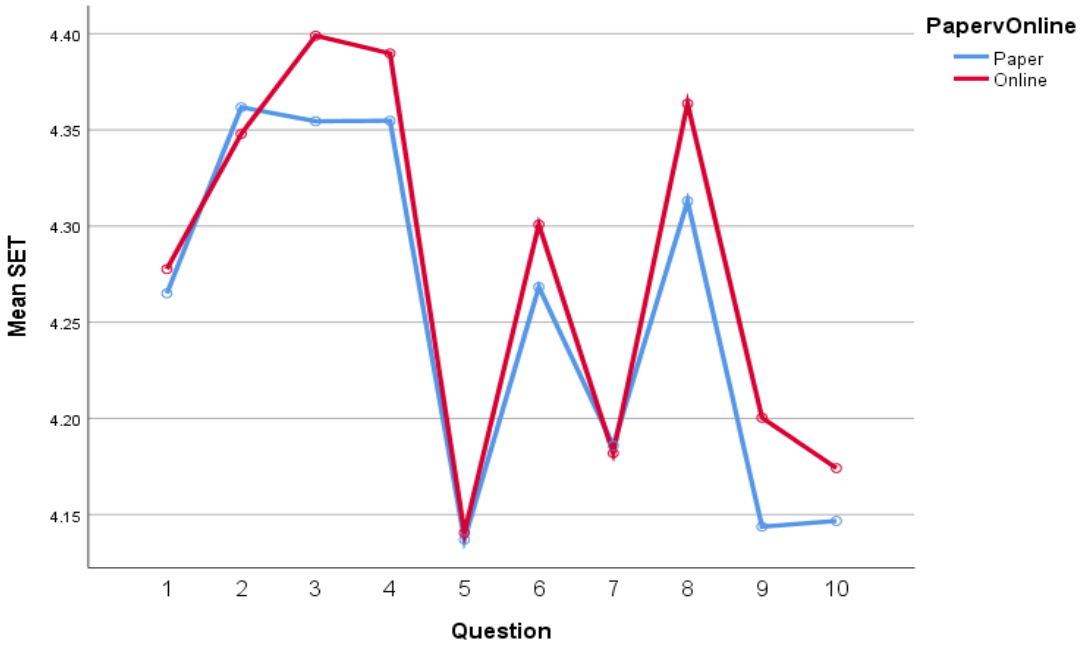

Figure 4.4

Graph for SET Rating by Data Collection Method

While the online data collection method was slightly higher for some question means, the data collection method accounted for no more than one tenths of a percent spread between the data collection methods. Question two (The instructor was well organized and prepared for class) was the only question where pencil and paper collection method had a higher mean.

CHAPTER 5

DISCUSSION AND CONCLUSIONS

The purpose of the present study was to investigate whether the factors of instructor variables, course variables, and SET administration modality, shown to be biasing in the literature, are also significant in a business school housed within a large, urban, public, R1 institution. A significant, biasing effect of SET outcomes could not be found for instructor gender, class size, or data collection method; however, section-level GPA, (instructor) race, and quantitative nature of the course (content) were all found to have significantly biased SET outcomes.

Major Findings

Instructor variables of gender and race

Although gender bias is prevalent in the business world, in educational roles, male instructors tend to be regarded as "professors" and women as "teachers" (Miller & Chamberlin, 2000). Nonetheless, gender bias was not found in this study, confirming numerous other studies where gender differences in SET outcomes were not found (Aleamoni, 1999; Basow & Distenfeld, 1985; Bennet, as cited in Driscoll et al., 2014, p. 294; Feldman, 1983; Feldman, 1992; Feldman 1993; Goodwin & Stevens, 1993; Hancock, Shannon, & Trentham, 1993). While this study posited results might be different in a business school and reflect the gender bias prevalent in the field of business, this study confirmed the findings of other studies (Feldman, 1993; Tatro, 1995) that found higher scores for female instructors on SETs.

This study explored a gap in the literature where only a paucity of research has focused specifically on African American faculty (Hawkins & Smith, 2011) and therefore

its findings are especially valuable since business schools, along with fields like engineering and math, are Caucasian dominated. There were not enough Latinx instructors to include them as a separate group in this study. This study found a significant difference as a function of race which accounted for 5.6 percent of the variance in SET outcomes. White instructors had the highest mean SET scores on the composite mean, while Asian instructors had the lowest. However, differences among racial groups varied from question to question and included a substantial finding: for questions five (Instructor feedback was useful) and eight (Course content was consistent with educational objectives) there was almost no observable variance between instructors of African American, Asian, and Other racial groups. While White instructors score higher on all questions, differences among other racial groups varied at the question level with some questions suggesting racial bias and other questions converging close to the same data point.

Course variables of class size and content

Bivariate correlations between class size and SET outcomes revealed class size does not have a meaningful effect on SET outcomes on individual questions or the composite mean, accounting for less than 1 percent of the variance. Large class sizes can decrease student-faculty communication, both in and out of the classroom, and affect student-faculty relationships (Bettinger & Long, 2018), and several previous studies found larger class sizes had a negative impact on SET outcomes (Bedard & Kuhn, 2008; Mandel & Sussmuth, 2008; McCahan & Rolheiser, 2018). While there are several reasons why larger class sizes can negatively impact SET outcomes – such as not being able to learn student names, decreased faculty-to-student individual communication and

itemized personal feedback, decreased student engagement and participation, and less active learning opportunities – several studies could not replicate these findings (Aleamoni & Graham, 1977; McPherson, 2006; McPherson et al, 2009; Ting, 2000). This study found class size had no meaningful effect on SET outcomes, specifically within a large, urban, public, R1 institution.

Instructors teaching quantitative courses are negatively affected in their SET outcomes, based solely on the quantitative nature of the business course(s) they teach. This study found quantitative courses have significantly lower means than qualitative courses on all SET questions and the composite mean, with content accounting for 2.5 percent of the variance. For all questions, the courses labeled as "mixed" have a mean between the quantitative and qualitative courses, further supporting the conclusion that level of quantitative vs qualitative content affects SET outcomes. This finding supports the literature that teaching quantitatively dominant courses, such as natural sciences and STEM, negatively affects SET outcomes.

Student variables of section-level GPA and SET

This study found section-level GPA had statistically significant, positive relationships with all 10 questions and the composite mean, which accounted for 11.1 percent of the variance. The essential conclusion being, with the section as the unit of analysis, higher SET outcomes are correlated with higher GPAs at the section level. This finding fits within the existing literature concluding expected grade correlates positively with SET outcomes.

Data collection method: paper and pencil vs online

While previous studies have focused on response rates and anonymity in data

collection, no previous study has systematically examined whether moving from paper administration to online has additional effects, specifically in a business school housed within a large, urban, public, R1 institution. Although there was a statistically significant difference between paper and pencil and online data collection methods, this was due to the large sample size and the data collection method accounted for no more than .10%. Essentially, the data collection method did not have an observable effect on SET outcomes. Since most survey and customer satisfaction research similar to SET matrix questions are now completed online, and student respondents clearly report higher satisfaction with online survey completion (Layne, DeCristoforo, & McGinty, 1999; Donovan, Mader & Shinsky, 2007) no further research is suggested. Even if there was an observable difference, the overwhelming time and financial resources saved through an online data collection method outweighs any differences. Instead of continuing to use pencil and paper SETs, the effects of online data collection should be mitigated as best as possible.

Limitations of the Study

One limitation of the study is the use of section level-data as the unit of analysis. Although instructor-level demographic data were used for the 5,701 sections included in the study, no demographic data at the student-level were included. Therefore, the research questions involving race and gender research were limited to instructor characteristics. And while one of three instructors in the study was female (33.3%), and therefore representing a large sample size (n=182 instructors, n=1901 sections), there were not enough data points to include Latinx (2%) and American Indian/Alaskan Native (0.4%) in the study (they were aggregated into Other), and only 5.5% of the population

was African American.

Another limitation of this study was that it only included data from the business school housed within a large, urban, public, R1 institution. The decision to limit the study to a school of business was by design in order to reduce differences in SET outcomes across academic disciplines and to ensure consistency with the extant literature. Although the findings of this study are generally congruent with the findings of research conducted at the university-level of the participating institution, this study lacks external validity. Therefore, the conclusions of this study may not generalize to other academic schools and colleges within the participating institution or to other universities. Additionally, the dataset is from 2009-2015; while this was helpful is studying the effect of the data collection method, three years of pencil and paper from 2009-2012, three years of online collection 2012-2015, it is nonetheless five-years-old and newer data from 2015-2020 could be used to retest some of the research questions.

Recommendations for Future Research

Further research regarding the racial biasing nature of SET outcomes for specific racial groups of instructors such as African Americans, Latinx, Indians, and Asians is needed. Specifically, additional research should focus on why certain questions are more biased by race than others. Also of primarty importance is to study the impact of SET bias on Instructor careers. Ways to combat found biases are needed.

Future research within the business school should include additional instructor and SET outcome data from 2016-2020 and focus on whether class size affects section-level GPAs. Future research should also focus on whether larger class sizes affect student learning outcomes as measured by common course assessments. Additionally, as within

this study, future research should control for Withdraws and Incompletes by using the final grades given as n= for class size, not the drop/add metric. Students' reasoning why the course was selected, such as required vs elective, and students' interest level in the subject matter are additional variables for future research.

Another area for future research is why students choose to answer some SET questions while skipping others. This study found most class sections do not have a consistent response rate across SET questions. Future research should focus on what types of questions students chose to not answer and their reasoning behind those decisions.

Incentives to Increase Response Rate of Online SETs

SET outcomes are only as useful as they are representative of the perceptions of students and the alignment between expectations and satisfaction. Therefore, response rate is of critical importance in capturing these important measures of students' experiences and perceptions. As an incentive to increase response rates, students are offered flexibility and choice through the sharing of results of specific, previous courses and sections' SET question outcomes. These four questions included two about feedback (Instructor provided useful feedback about exams, projects, and assignments) and equitable grading (Instructor has applied grading policies fairly), and the two administration weighed heavily in promotion, tenure, and merit decisions, questions seven (Instructor taught well) and 10 (I learned a great deal) heavily in decision-making. Future research should focus on the effects of sharing SET outcomes at the time of course registration, to assist in instructor and course selection. While the business school explicitly states assisting students in instructor selection is not the objective of SET

outcomes sharing, it is hard to imagine any other outcome – intended or otherwise.

Student feedback is an essential part of faculty professional development and improved course design, and high response rates are important in eliminating non-response bias. While offering extra credit for SET completion in a quid pro quo manner is obviously unethical and would likely result in obvious SET outcome inflation, Scriven (1981) argued "that it is unethical to deny students the opportunity to view the results of ratings which they have engendered" (as cited by Watchel, 1998, p. 205). Further research is needed to settle this debate; however, universities and instructors across the industry typically use more subtle approaches for incentives to complete SETs such as allowing time in class to complete SETs and withholding final grades.

More balanced approach to measuring teaching and learning

In conjunction with existing literature, results of this study prove that although SETs can be biased in certain contexts with certain demographics, the bias can be situational, needs to be delineated, and generalizations should be made with caution.

As shown with race, the actual SET questions matter, and certain questions can reveal more bias that others. During the time of this study, the business school weighed questions seven (The instructor taught well) and 10 (I learned a great deal) very heavily in merit, tenure and promotion-related decision-making. Data included in this six-year study were collected using a 14-question framework with 11 Likert scale questions ("I came well-prepared for class" was not included, as this item was deemed as not measuring teaching) and three open-ended questions. These 14 questions were mandated across the university for all single instructor lecture model sections. The model has since been recently changed to add flexibility into the model and more choice across the

administrative organizational chart. While the three open-ended questions remained the same, the university mandated questions were limited to four.

Table 5.1

Changes in University-Wide SET Questions

New Question	Change
The instructor was organized and prepared for class.	Removed "well" from "well organized."
So far, the instructor has applied grading policies fairly.	No change.
Overall, the instructor was effective in helping me learn the material in this course.	Formerly "The instructor taught this course well."
Overall, I learned a great deal from this course.	Added the word "Overall."

From there, flexibility and choice are exercised by choosing questions from a bank of over 150 items (1) at the school-level where each college chooses three additional questions to be used school-wide, (2) at the department-level where each department chooses three additional questions to be used department-wide, (3) and at the instructor-level where each instructor can choose up to four questions additional questions to be used in each of their sections. Additionally, two questions may be added to General Education, Writing Intensive, Honors, and Online courses. The additional flexibility and choice allows individual colleges, administrators, departments, and instructors, to have a voice and role in the SET process.

Conclusion

In short, although the SET form reveals bias in some areas and not others, it is the overemphasis placed on the SET outcomes that negatively affect personnel decisions. A more balanced approach to measuring teaching effectiveness that includes peer review, professional growth, and academic student outcomes need to be included in the decision-making processes, in conjunction with SET outcomes.

REFERENCES

AACSB. (2015). AACSB 2014-2015 Deans Survey. Tampa, Florida: AACSB.

AACSB. (2016). Gender Equality and Pay Gap in Business Education: EMEA Region (Part I). Tampa, Florida: AACSB. Retrieved 2016, from http://aacsbblogs.typepad.com/dataandresearch/gender/

AACSB. (2019). Business School Data Guide. 2019 AACSB International.

Abbott, R. D., Wulff, D. H., Nyquist, J. D., Ropp, V. A., & Hess, C. W. (1990). Satisfaction With Processes of Collecting Student Opinions About Instruction: The Student Perspective. Journal of Educational Psychology, 82(2), 201-206.

Abdulla, M., Badri, M., Dodeen, H., & Kamali, M. A. (2006). Identifying potential biasing variables in student evaluation of teaching in a newly accredited business program in the UAEaccredited business program in the UAE. International Journal of Educational Management, 20(1), 43-59.

Abrami, P., d'Apollonia, S., & Rosenfield, S. (2007). Abrami, P. C., d'Apollonia, S., & Rosenfield,. In r. P. Perry, & J. C. Smart, The scholarship of teaching and (pp. pp. 385–445). Dordrecht, Netherlands: Springer.

Aleamoni, L. M. (1999). Student Rating Myths Versus Research Facts From 1924 to 1998. Journal of Personnel Evaluation in Education, 13(2), 153–66.

Aleamoni, L., & Graham, M. (1977). The Relationship Between CEQ Ratings and Instructor Rank, Class Size, and Course Level. Journal of Educational Management, 11(3), 189-202.

Anderson, K., & Smith, G. (2005, May). Students' Preconceptions of Professors: Benefits and Barriers According to Ethnicity and Gender. Hispanic Journal of Behavioral Sciences, 27(2), 184-201.

Arreola, R. (2006). Developing a Comprehensive Faculty Evaluation System: A Guide to Designing, Building, and Operating Large-Scale Faculty Evaluation Systems. San Francisco, CA: Jossey-Bass.

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic Course Evaluations: Does an Online Delivery System Influence Student Evaluations? The Journal of Economic Education, 37(1), 21-37.

Basow, S. A. (1995). Student Evaluations of College Professors: When Gender Matters. Journal of Educational Psychology, 656-665.

Basow, S. A. (2000). Best and worst professors: Gender patterns in students' choices. Sex, 34, 407-417, Sex.

Basow, S. A., & Distenfeld, S. M. (1985). Teacher expressiveness: More important for males than females? Journal of Educational Psychology, 77, 45-52.

Basow, S., & Martin, J. (2012). Bias in Student Evaluation. In M. Kite, Effective evaluation of teaching: A guide for faculty and administrators (pp. 40-49). Society for the Teaching of Psychology.

Basow, S., & Montgomery, S. (2006). Student Ratings and Professor Self-Ratings of College Teaching: Effects of Gender and Divisional Affiliation. Journal of Personnel Evaluation in Education , 18, 91–106.

Basow, S., Martin, J., & Codos, S. (2013, September 14). The Effects of Professors' Race and Gender on Student Evaluations and Performance.

Baum, P., Chapman, K., Dommeyer, C., & Hanna, R. (2001). Online versus in class student evaluations of faculty. Paper presented at the Hawaii Conference on Business. Honolulu.

Bavishi, A., Helb, M., & Madera, J. (2010). The Effect of Professor Ethnicity and Gender on Student Evaluations: Judged Before Met. Journal of Diversity in Higher Education, 3(4), 245-256.

Bedard, K., & Kuhn, P. (2008). Where class size really matters: Class size and student ratings of instructor effectiveness. Economics of Education Review, 27, 253–265.

Bennet, S. (1982). Student Perceptions of and Expectations for Male. Journal of Educational Psychology, Vol. 74, No. 2, 170-179.

Bettinger, E., & Long, B. (2018). Mass Instruction or Higher Learning? the Impact of College Class Size on Student Retention and Graduation. Education Finance and Policy, 13(1), 97-118.

Blunt, A. (1991). The effects of anonymity and manipulated grades on student ratings of instructors. Community College Review, 18(4), 48-54.

Bonitz, V. S. (2011). Student evaluation of teaching: Individual differences and bias effects. ProQuest LLC. Ann Arbor, MI, USA: ProQuest LLC.

Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching. ScienceOpen, 1-11.

Brandenburg, G. C., & Remmers, H. H. (1927). A rating scale for instructors. Educational Administration and Supervision, 13, 399-406.

Carter, J. (2018, May 24th). USC nixes student evaluations as part of tenure review. Retrieved from Education Dive: https://www.educationdive.com/news/usc-nixes-student-evaluations-as-part-of-tenure-review/524163/

Cashin, W. (1990). Students do rate different academic fields differently. New Directions for Teaching and Learning, 43, 113-121.

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

Chan, C., Luk, L., & Zeng, M. (2014). Teachers' perceptions of student evaluations of teaching. Educational Research and Evaluation, 20(4), 275-289.

Cobanoglu, C., & Cobanoglu, N. (2003). The effect of incentives in web surveys: application and ethical considerations. International Journal of Market Research, 45(4), 475-488.

Crawford, S. D., Couper, M. P., & Lamias, M. J. (2001, Summer). Web Surveys: Perceptions of Burden. Social Science Computer Review, 19(2), 146-162.

Davis, B. G. (2009). Tools for Teaching. San Francisco, CA: Jossey-Bass.

DiPietro, M., & Faye, A. (2005). Online student-ratings-of-instruction (SRI) mechanisms for maximal feedback to instructors. The 30th Annual Meeting of the Professional and Organizational Development Network. Milwaukee.

Doerer, K. (2019, January 13th). Colleges Are Getting Smarter About Student Evaluations. Here's How. Retrieved from the Chronicle of Higher Education: https://www.chronicle.com/article/Colleges-Are-Getting-Smarter/245457

Dommeyer, C. J., Baum, P., & Hanna, R. W. (2002). College Students' Attitudes Toward Methods of Collecting Teaching Evaluations: In-Class Versus On-Line. Journal of Education for Business, 78(1), 11-15.

Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004, October). Gathering faculty teaching evaluations by in-class and online surveys: their effects on response rates and evaluations. Assessment & Evaluation in Higher Education, 29(5), 611-623.

Donovan, J., Mader, C. E., & Shinsky, J. (2006, Winter). Constructive student feedback: Online vs. traditional course evaluations. Journal of Interactive Online Learning, 5(3), 283-296.

Donovan, J., Mader, C., & Shinsky, J. (2007, Winter). Online vs. Traditional Course Evaluation Formats: Student Perceptions. Journal of Interactive Online Learning, 6(3), 158-180.

Driscoll, A., Hunt, A. N., & MacNell, L. (2014, December). What's in a Name: Exposing Gender Bias in Student. Innovative Higher Education.

Durden, G., & Ellis, L. (1995). The Effects of Attendance on Student Learning in Principles of Economics. Papers and Proceedings of the Hundredth and Seventh Annual Meeting of the American Economic Association. 85, pp. 343-346. Washington D.C.: American Economic Association.

Elson, R., Gupta, S., & Krispin, J. (2018). Students' perceptions of instructor interaction, feedback, and course. Journal of Instructional Pedagogies effectiveness in a large class environment, 20, 1-19.

Estelami, H. (2015). The Effects of Survey Timing on Student Evaluation of Teaching Measures Obtained Using Online Surveys. Journal of Marketing Education, 37(1), 54-64.

Feldman, K. (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. (1976). Grades and college students' evaluations of their courses and teachers. Research in Higher Education,, 6, 69-111.

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. (1992). College students' views of male and female college teachers: part 1. Research in Higher Education , 317-318.

Feldman, K. A. (1993). College students' views of male and female college teachers: Part II . Research in Higher Education, Volume 34 (2), 151-211.

Fike, D. S., Doyle, D. J., & Connelly, R. J. (2010). Online vs. Paper Evaluations of Faculty: When Less is Just as Good. The Journal of Effective Teaching, 10(2), 42-54.

Foschi, M. (2000). Double Standards for Competence: Theory and Research. Annual Reviews of Sociology, 26, 21-42.

Freishtat, R., & Stark, P. (2014). An Evaluation of Course Evaluations. ScienceOpen.

Glass, G., & Smith, M. (1979, January-February). Meta-Analysis of Research on Class Size and Achievement. Educational Evaluation and Policy Analysis, 1(1), 2-16.

Gorry, D. (2017). The impact of grade ceilings on student grades and course evaluations: Evidence from a policy change. Economics of Education Review, 56, 133-140.

Goyder, J. (2001). The Silent Majority: Non-Respondents on Sample surveys. In R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, & J. Waksberg, Telephone Survey Methodology. New York: Polity Press.

Griffin, K., & Reddick, R. (2011, October). Surveillance and Sacrifice: Gender Differences in the Mentoring Patterns of Black Professors at Predominantly White Research Universities. American Educational Research Journal, 48(6), 1032-1057.

Griffin, T., Hilton, J., Plummer, K., & Barret, D. (2014). Correlation between grade point averages and student evaluation of teaching scores: taking a closer look. Assessment & Evaluation in Higher Education, 39(3), 339-348.

Hamermesh, D., & Parker, A. (2005). Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity. Economics of Education Review, 24, 369-376.

Hancock, G. R., Shannon, D. M., & Trentham, L. L. (1993). Student and Teacher Gender in Ratings of University. Journal of Personnel Evaluation in Education, Volume 6, 235-248.

Harlow, R. (2003). Race Doesn't Matter, but: The Effect of Race on Professors' Experiences and Emotion Management in the Undergraduate College Classroom. Social Psychology Quarterly, 66(4), 348-363.

Hawkins, B., & Smith, B. (2011). Examining Student Evaluations of Black College Faculty : Does Race Matter? The Journal of Negro Educatio, 80(2), 149-162.

Howard, G., & Maxwell, S. (1982). Do Grades Contaminate Student Evaluations of Instruction? Research in Higher Education, 16(2), 175-188.

Kherfi, S. (2011). Whose Opinion Is It Anyway? Determinants. The Journal of Economic Education of Participation in Student Evaluation of Teaching, 42(1), 19-30.

Latta, R., & Warren, D. (1978). Individual Differences Model Applied to Instruction and Evaluation of Large College Classes. Journal of Educational Psychology, 70(6), 960-970.

Layne, B. H., DeCristoforo, J. R., & McGinty, D. (1999). Electronic Versus Traditional Student Ratings of Instruction. Research in Higher Education, 40(2), 221-232.

Littleford, L., Ong, K., Tseng, A., Milliken, J., & Humy, S. (2010). Perceptions of European American and African American Instructors Teaching Race-Focused Courses. Journal of Diversity in Higher Education, 3(4), 230-244.

Mandel, P., & Sussmuth, B. (2008). The Relevance and Hicksian Surplus of Preferred College Class Size. Leipzig, Germany: Institute for Empirical Research in Economics.

Marsh, H. (1980). The Influence of Student, Course, and Instructor Characteristics in Evaluations of University Teaching. American Educational Research Journal, 17(1), 219-237.

Marsh, H. (1987). Students' Evaluation of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research. International Journal of Educational Research, 11, 253-388.

Marsh, H. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. Perry, & J. Smart, The

Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective (pp. 319–383). Springer.

Marsh, H., & Overall, J. U. (1981). The Relative Influence of Course Load, Course Type, and Instructor on Student? Evaluations of College Teaching. American Educational Research Journal, 18(1), 103-112.

McCahan, S., & Rolheiser, C. (2018). University of Toronto's Cascaded Course Evaluation Framework: Validation Study of the Institutional Composite Mean (ICM). Toronto, ON: The Centre for Teaching Support & Innovation.

McKeachie, W. J. (1979, October). Student Ratings of Faculty: A Reprise. Academe, 65(6), 384-397.

McKeachie, W. J., & Svinicki, M. D. (2006). McKeachie's teaching tips: strategies, research, and theory for college and university teachers. Boston: Houghton Mifflin.

McPherson. (2006, February). Determinants of How Students Evaluate Teachers. The Journal of Economic Education, 3-20.

McPherson, M. A., Jewell, R. T., & Kim, M. (2009). What Determines Student Evaluation Scores? A Random Effects Analysis of Undergraduate Economics Classes. Eastern Economic Journal, 35, 37–51.

Miller, J., & Chamberlin, M. (2000). Women are teachers, men are professors: A study of student perceptions. Teaching Sociology, Volume 28, pages 283-298.

Murray, H. G. (2007). Low-inference Teaching Behaviors and College Teaching. In R. P. Perry, & J. C. Smart, The Scholarship of Teaching and Learning in Higher

Education: an evidence-based perspective (pp. 184-200). Dordrecht, Netherlands: Springer.

Noland, M., Moran, T., & Kotschwar, B. (2016). Is Gender Diversity Profitable? Evidence from a Global Survey. Washington, DC: Peterson Institute for International Economics.

Ory, j. (2001). Faculty Thoughts and Concerns About Student Ratings. New Directions for Teaching and Learning, 87, 3-15.

Overall, J. U., & Marsh, H. W. (1982, December). Students' Evaluations of Teaching: An Update. American Association for Higher Education Bulletin, 9-12.

Perry, A., Wallace, S., Moore, S., & Perry-Burney, G. (2015, November 16th). Understanding Student Evaluations: A Black Faculty Perspective. Reflections: Narratives of Professional Helping, 20(1), 29-35.

Porter, S. R. (2004). Raising Response Rates: What Works? NEW DIRECTIONS FOR INSTITUTIONAL RESEARCH, 121, 5-12.

Provost, O. o. (2019, August). Office of the Provost. Retrieved from University of Oregon : https://provost.uoregon.edu/revising-uos-teaching-evaluations

Pulich, M. A. (1984). Ratings: Better Use of Student Evaluations for Teaching Effectiveness. Improving College and University Teaching, 32(2), 91-94.

Ramsden, P. (1991). A Performance Indicator of Teaching Quality in Higher Education: the Course Experience Questionnaire. Studies in Higher Education, 16(2), 129-150.

Reisenwitz, T. (2016). Student Evaluation of Teaching: An Investigation of Nonresponse Bias in an Online Context. Journal of Marketing Education, 38(1), 7-17.

Rosenfeld, P., Booth-Kewley, S., & Edwards, J. E. (1993, March). Computer-Administered Surveys in Organizational Settings: "Alternatives, Advantages, and Applications". The American Behavioral Scientist, 36(4), 485-511.

Rush, L., Shaw, D., & Young, S. (2009). Evaluating Gender Bias in Ratings of University Instructors' Teaching Effectiveness. International Journal for the Scholarship of Teaching and Learning, 3(2, article 19).

Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003, August). Assessing Response Rates and Nonresponse Bias in Web and Paper Surveys. Research in Higher Education, 44(4).

Scheck, C., Kinicke, A., & Webster, J. (1994). The Effect of Class Size on Student Performance: Development and Assessment of a Process Model. Journal of Education for Business, 70(2), 104-111.

Scriven, M. (1981). Summative Teacher Evaluation. In J. Millman, Handbook of Teacher (pp. 244-271). Beverly Hills, California: Sage.

Sinclair, L., & Kunda, Z. (2000). Motivated Stereotyping of Women: She's Fine if She Praised Me but Incompetent if She Criticized Me. Personality and Social Psychology Bulletin, 1329-1342.

Spooren, P., Brockx, B., & Mortelmans, D. (2013, December). On the Validity of Student Evaluation of Teaching: The State of the Art. Review of Educational Research, 83(4), 598-642. doi:10.3102/0034654313496870

Sproull, L., & Kiesler, S. (1991). Connections: New Ways of Working in the Networked Organization. Cambridge, MA, USA: MIT Press.

Stewart, R. J., & Phelps, R. E. (2000). Faculty of color and university students: Rethinking the evaluation of faculty teaching. Journal of the Research Association of Minority Professors, 4, 49-56.

Stroebe, W. (2017). Why Good Teaching Evaluations May Reward Bad Teaching: On Grade Inflation and Other Unintended Consequences of Student Evaluations. Perspectives on Psychological Science, 11(6), 800-816.

Tatro, C. N. (1995). Gender effects on student evaluations of faculty. Journal of Research and Development in Education, 28, pp. 169-173.

Ting, K.-f. (2000). A Multilevel Perspective on Student Ratings of Instruction: Lessons from the Chinese Experience. Research in Higher Education, 41(5), 637-661.

Uttl, B., White, C., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. Studies in Educational Evaluation, 54, 22-42.

Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. PeerJ.

Wachtel, H. K. (1998). Student Evaluation of College Teaching Effectiveness: a brief review. Assessment & Evaluation in Higher Education, 23(2), 191-212.

White, G. (2017, October 26). There Are Currently 4 Black CEOs in the Fortune 500. The Atlantic.

Williams, D. (2007). Examining the Relation between Race and Student Evaluations of Faculty Members: A Literature Review. Profession, 168-173.