2022

A Multi-Method Examination of the Effects of Students' Unconscious Biases on Student Evaluations of Instructors

Brittany M. Kowalski
West Virginia University, bmkowalski@mix.wvu.edu

A Multi-Method Examination of the Effects of Students' Unconscious Biases on Student Evaluations of Instructors

Brittany M. Kowalski

Dissertation submitted
to the Eberly College of Arts and Sciences
at West Virginia University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in
Sociology

Lisa M. Dilks, Ph.D., Chair
Melissa Latimer, Ph.D.
Rachel E. Stein, Ph.D.
Sharon R. Bird, Ph.D.

Department of Sociology and Anthropology

Morgantown, West Virginia
2022

Keywords: roles, role congruity, gender, gender roles, student evaluations, race/ethnicity

ABSTRACT

A Multi-Method Examination of the Effects of Students' Unconscious Biases on Student Evaluations of Instructors

Brittany M. Kowalski

In this dissertation, I complete three studies to evaluate potential reactions to target role congruity, especially gender role congruity, through an examination of Student Evaluations of Instructors (SEIs). Target role congruity refers to assessments an observer makes of whether or not the various roles a target person fills "fit" with one another. For example, a woman surgeon may be perceived as being in an incongruent role due to the masculine characteristics associated with the occupation and the continued dominance of men in the field. Researchers utilizing congruity theories has shown that both women and men in roles that are incongruent to their gender are viewed as less competent and less acceptable than those whose traits conform to their gender roles. People in gender role incongruity roles tend to receive sanctions and backlash commonly exhibited through negative evaluations due to their perceived role incongruity.

In these studies, I examine how target role congruity affects a particular type of subjective evaluation, student evaluations of instruction (SEIs). In the first two studies, I use exploratory factor analysis, confirmatory factor analyses, multiple-indicators and multiple-causes (MIMIC) models, and grouped structural equation models (SEM) to evaluate how instructor role congruity may affect quantitative SEI measures. In Study 1 (Chapter 2), I determine whether the questions that comprise the SEI forms are biased in measurement depending upon the role congruity of the instructor as determined by their gender and the discipline in which they teach. In Study 2 (Chapter 3), I extend these results by examining whether the race/ethnicity of the instructor moderates the effect of gender role congruity on quantitative SEI measures. Finally, in Study 3 (Chapter 4), I complete a qualitative analysis of open-response SEI questions in order to determine the potential causal mechanisms behind any differences in SEI scores by instructor gender, level of gender role congruity, and race/ethnicity.

The results of Study 1 (Chapter 2) indicate that when measurement invariance is accounted for, differences in SEI scores based on instructor gender and level of gender role congruity are eliminated. The results of Study 2 (Chapter 3) indicate that when measurement invariance is accounted for, some differences in SEI scores based on instructor gender, level of gender role congruity, and race/ethnicity are eliminated while some persist. These results indicate the importance of measurement invariance testing as well as the importance of considering instructor role congruity when examining SEI scores. Study 3 (Chapter 4) results further indicate that the various statuses of instructors may influence how students perceive and evaluate their courses.

Future research using congruity theories should consider how other salient social roles may moderate the effects of perceptions of target role congruity on subjective evaluations. Additionally, future research on student evaluations should consider the inclusion of more instructor statuses as well as other potential mitigating factors such as student statuses and course characteristics in their evaluations.

DEDICATION

To my family and friends who have stood by my side throughout the years. To my Grandpa Al, Grandpa Marion, and Grandma Gigi, I wish you were here to celebrate and I hope I have made you all proud. To my Grandma Cheri, thank you for always being there and celebrating every victory with me. To my parents, Karen and Greg, thank you for setting me on the path that brought me here and never giving up on me. To my husband Jamie, thank you for always being a shoulder to cry on, an ear to listen, and a smile to brighten my day. To my Rat Pack, you helped me become the person I am and have cheered me on through every step of this journey, without you I would not be half the woman I am today. I would not be here completing my doctorate if it were not for each and every one of you and I am eternally grateful for your never-ending love and support. Na Zdrowie!

ACKNOWLEDGMENTS

I first took a sociology class in high school many years ago. Never did teenage Brittany think she would one day be writing the acknowledgments for her dissertation to complete a Ph.D. in the discipline, but the day has finally come. I would be remiss if I did not acknowledge the brilliant professors who helped me on this journey.

To Dr. Leslie Bunnage who encouraged me to pursue graduate school and was always there for me when I randomly appeared at her office door, thank you. Thank you for taking a confused college freshman who was not sure what she wanted and helping her to become a sociologist. Thank you for all of the laughs and memories, I hope I have made you and all of the sociology faculty at Seton Hall proud.

To the faculty in the Department of Sociology and Anthropology at West Virginia University, thank you all for your teaching, scholarship, and guidance over the years. I had the privilege of working closely with so many of you as a student, teaching assistant, and co-author. I do appreciate every lesson you all taught me through both formal classes and informal interactions. Thank you all for your dedication to sociology and your students, your effort does not go unnoticed.

To my committee members, Dr. Rachel Stein, Dr. Sharon Bird, and Dr. Melissa Latimer, thank you all for your thoughtful comments and feedback throughout the years. I am a better scholar, researcher, and writer thanks to each and every one of you. Additionally, to Dr. Rachel Stein who took a chance and hired me as a research assistant after my first year of graduate school. I never thought studying Amish communities could be so fascinating, but your brilliance and sociological perspective made it an incredible learning and personal experience. Furthermore, to Dr. Sharon Bird, thank you for the years of working together. I learned so much about navigating research and networking because of you. Thank you for all that you taught me and helping me become a more thoughtful scholar.

Finally, to my dear committee chair Dr. Lisa Dilks, thank you for believing in me and pushing me to become a stronger scholar and person. No matter the number of drafts or little detail-oriented questions I had, you were always there for me providing helpful feedback and ready to bounce ideas back and forth. I would not be here today if it were not for you hearing the idea for my thesis proposal many years ago and knowing then and there that we needed to work together. I have enjoyed our working and personal relationships so much and I would not be where I am today without you. I look forward to calling you a friend for many years to come.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Student Evaluation of Instruction	SEI
Science, Technology, Engineering, and Math	STEM
Social Role Theory	SRT
Role Congruity Theory	RCT
Status Incongruity Hypothesis	SIH
Exploratory Factor Analysis	EFA
Confirmatory Factor Analysis	CFA
Multiple Indicators Multiple Causes	MIMIC
Structural Equation Model	SEM

CHAPTER 1: INTRODUCTION

INTRODUCTION

A subjective evaluation is an assessment of something that is highly influenced by an evaluator's feelings and is opinionated, potentially biased, and not based upon measurable criteria. Subjective evaluations are used on an almost daily basis to communicate preferences and impressions in both informal and formal ways such as a friend describing their experience at a new restaurant or a review of the restaurant from a food critic. Both reviews are subjective in that they filtered through the impressions of a person and not necessarily based upon objective measures even though one review is from an expert and the other is not.

Research has shown many factors can affect subjective evaluations with stereotypes such as those associated with gender, race, and age, being a major source of bias in subjective evaluations (Arbuckle and Williams 2003; Kobrynowicz and Biernat 1997; Liden, Stilwell, and Ferris 1996; Smith et al. 2001; Smith et al. 2019). Characteristics or statuses such as gender, race, and age are categories of classifying persons that tend to be highly salient and memorable, more so than other characteristics like eye color, hair color, clothing, or names because they are associated with behavioral and attitudinal expectations (Burn 1996). For example, with respect to gender there are societal expectations, known as gender roles[1], for the behaviors, attitudes, and beliefs of men that are different for women. Men are traditionally expected to be breadwinners while women are traditionally expected to be caretakers. Gender role expectations are activated

[1] While some researchers refer to the social positions a person holds as a role, others refer to them as a status. At their core, the words role and status both refer to a social position a person holds which carries expectations for behaviors, attitudes, and beliefs and places a person in a social hierarchy relative to others due to their roles/statuses. The expectations of a role/status are activated when a person perceives a target to be a member of a particular role or status group. The processing of a target into a particular role/status then leads to automatic assumptions about the person's behaviors, attitudes, and beliefs due to the role/status they are perceived to belong to. In this dissertation, both words are used to denote a socially held position that is relative to other social positions and carries behavioral, attitudinal, and belief expectations. Any reactions that occur as a result of a target's role/status are due to the perceptions of that role and its expectations and are not caused by the target themselves.

almost immediately upon meeting a person and identifying the person's gender group and then influence interactions and future expectations from that moment forward (Eagly, Wood, and Diekman 2000; Heilman 2012a).

Research has repeatedly shown that a person's belonging to particular status groups and the role expectations associated with those groups affects the subjective evaluations they receive from others (e.g. Arbuckle and Williams 2003; Kobrynowicz and Biernat 1997; Smith et al. 2019). Even though they can be biased and influenced by a target persons' characteristics and the stereotypes and roles associated with those characteristics, subjective evaluations are still relied upon heavily and therefore can have a profound impact on the lives of those being evaluated. For example, subjective evaluations can affect employment decisions such as who is hired, who is retained, and who is promoted (Foschi, Lai, and Sigerson 1994; Liden et al. 1996; Smith et al. 2001, 2019).

In this dissertation, I examine another potential source of bias in the content and completion of subjective evaluations—reactions to target role congruity. Target role congruity refers to assessments an observer makes of whether or not the various roles a target person fills "fit" with one another. For example, a woman surgeon may be perceived as being in an incongruent role due to the masculine characteristics associated with the occupation of surgeon and the continued man dominance in the field of medicine. In this dissertation, I specifically examine perceptions of role (in)congruity as a source of bias in both quantitative and qualitative subjective evaluations. I seek to answer the following questions: (1) Are subjective evaluations affected by perceptions of target role (in)congruity? (2) Does the effect of perceptions of target role (in)congruity also depend on the race/ethnicity of the target?

Researchers have found that target role congruity can affect subjective evaluations in a variety of areas such as dating (Hitsch, Hortaçsu, and Ariely 2010), employment potential (Foschi et al. 1994), and leadership ability (Garcia-Retamero and López-Zafra 2006; Smith et al. 2019). When a person is perceived to be gender role incongruent they are more likely to receive lower evaluations and/or other sanctions (Diekman and Eagly 2008; Fassiotto et al. 2018). For example, even highly competent women such as female physicians are rated lower than their male peers by medical residents (Fassiotto et al. 2018). I extend these and other findings by examining the effect of target role (in)congruity on a novel type of subjective evaluation: student evaluations of instruction.

Student evaluations of instruction (SEIs) are a type of subjective evaluation which has become almost universal in higher education in which students are asked to evaluate their instructors and courses at the end of each semester (Algozzine et al. 2004; Benton and Cashin 2014). Student evaluations have been shown to be heavily influenced by the gender and race/ethnicity of the instructor of the course (Boring, Ottoboni, and Stark 2016; Smith and Hawkins 2011). Because SEIs are frequently used in the consideration of a faculty member's retention, promotion, and tenure, any biases in student evaluations are highly problematic. One potential source of bias in student evaluations that has received considerably less attention is the gender role (in)congruity of the instructor with the discipline in which they teach (Basow 1995). If differences exist in student evaluations based on instructor gender and race/ethnicity and differences exist in subjective evaluations based on perceived target gender role incongruity, there is reason to believe that differences exist in student evaluations based not only on the gender and race/ethnicity of the instructor but also on the perceived level of gender role (in)congruity of the instructor with respect to the discipline in which they teach.

The purpose of this dissertation below is to analyze how SEIs as a particular type of subjective evaluation, may be affected by students' reactions to the perceived gender role (in)congruity of their instructors. I complete a three-part, multi-method examination of SEI data that includes faculty characteristics to examine the ways in which these factors affect the content of subjective evaluations of teaching. The analyses include two studies that consist of quantitative analyses of traditional close-ended SEI measures and a third study which consists of qualitative analyses of traditional open-ended SEI measures.

In the remainder of this chapter, I first discuss theories of gender role congruity. I then discuss some of the existing research on SEIs in order to describe the more specific context of the current studies followed by a brief summary of each of the analyses completed in each of the three studies. I conclude the chapter with an overview of the data used in all three studies. This chapter is followed by a chapter for each study (Chapters 2-4) and a concluding chapter (Chapter 5) in which the broader implications and future directions of this work are discussed.

OVERVIEW OF LITERATURE AND THEORY

This dissertation utilizes structural social psychological theories to consider how students' reactions to instructor gender role (in)congruity may affect the content of subjective evaluations of their instructors. A structural social psychological approach focuses on understanding how micro-level evaluations of instructors are affected by macro-level structures of inequality (Lawler, Ridgeway, and Markovsky 1993). Through this approach, I consider the effects of different statuses and their placement in hierarchies of gender and race on how students view and evaluate their instructors (Hollander and Howard 2000). Given that there is a historical pattern of White men being the overwhelming majority of instructors in higher education in general, but even more so in some fields such as science, technology, engineering,

and mathematics (STEM), it is crucial to understand how women and people of color who enter into these White-men dominated fields are perceived and evaluated given that they may be seen as incongruent to their occupational role.

Assumptions and derivations from Role Theory and Role Congruity Theory as well as Status Incongruity Hypothesis, another theory of congruity, serve as the core theoretical perspectives for this dissertation. In addition to these social psychological theories, my dissertation is heavily influenced by research on student evaluations as that is the context in which I examine the effects of gender role (in)congruity on subjective evaluations. I integrate key findings from the SEI literature with insights from social psychology and Role Theory to fully consider the diversity of methods and theories that have been applied to the study of student evaluations of instruction in higher education.

Role Theory and Social Role Theory

Role Theory is a social psychological perspective that considers how the roles people occupy and the expectations of those roles affect their own and others' behaviors, attitudes, and values (Jacobs 2018). Role expectations prescribe what a person in a particular role should do, how that person should behave, and what their attitudes and beliefs should be (Jacobs 2018). Roles happen at different levels such as the group, organization, or society, and can occur in both formal and informal situations at each level (Lynch 2007). People are recruited to societally available positions with each position dictating which activities the holder should partake in and how to interact with others (Jacobs 2018). Every person occupies multiple positions, or roles, and must learn the acceptable activities and actions of all of them (Jacobs 2018).

Roles carry with them prescriptions, expectations about what people actually do, and proscriptions, expectations about what a person ought to do or would ideally do, regarding their

behavior, values, and actions (Eagly et al. 2000). Role prototypes represent the ideal form of the role as determined by an individual's experiences and social statuses which serve as a barometer to which people who enact that role are held accountable (Johnson et al. 2008). Given agency, people cannot ever perfectly replicate the prescribed activities or ideal-types of social relationships as dictated by a role but rather individuals must approximately conform to the prescriptions to sufficiently occupy the role (Jacobs 2018).

Roles can be diffuse and specific with both types carrying norms and expectations with them. Diffuse roles are broad roles that occur in many/most situations that confer behavioral, attitudinal, and other expectations (Diekman and Schneider 2010). Gender, race, and age are examples of diffuse statuses because they are influential across most social situations to some degree (Koenig and Eagly 2014). Specific roles, such as occupation or parental status, are roles that occur in particular circumstances and do not cut across other roles (Diekman and Schneider 2010). People vary in the extent to which specific roles become a part of their identity, but the more internalized a specific role becomes, the more likely the person may be to carry out the beliefs and norms associated with that specific role (Diekman and Schneider 2010). For example, a very hands-on mom versus a more absent father would experience the specific role of parent role very differently (Diekman and Schneider 2010). The demands of specific roles, such as occupation, can affect the extent to which a person's behavior is determined by their diffuse roles, such as gender, and vice versa, a person's diffuse roles can affect their behavior in specific roles (Eagly et al. 2000).

Social Role Theory (SRT) extends Role Theory by focusing on the effect of role expectations within the social structure and how a person is affected both internally and externally when they occupy multiple roles, particularly when the norms of those roles are

conflicting (Eagly and Karau 2002). SRT posits that the differences in the observed social behaviors and personalities of men and women originate in the varying distribution of men and women into social roles (Eagly et al. 2000; Koenig and Eagly 2014). Social roles refer "to the shared expectations that apply to persons who occupy a certain social position or are members of a particular social category" (Eagly et al. 2000: 130). When a group is overrepresented relative to other groups in a social role, perceivers infer that the behaviors of that group are generalizable to everyone in the group and thus a group stereotype is born (Koenig and Eagly 2014). For example, women have been historically overrepresented in childcare, thus, the stereotype that women are warm, communal, and nurturing developed and persists (Koenig and Eagly 2014).

Correspondent inference is a psychological process that produces stereotypes of social groups that mirror the qualities they play out in their social roles (Eagly et al. 2000). People do not generally take the time to reason beyond the situation at hand and rather rely on their stereotype inferences to know how to act and what to expect from other actors in a given situation (Eagly et al. 2000). Masculine gender roles came to be typically associated with greater agency and competence than feminine gender roles which leads to masculinity being more compatible with leadership (Diekman and Eagly 2008). Feminine gender roles came to be more associated with communality and maintaining interpersonal relationships and are thus less associated with leadership (Diekman and Eagly 2008).

Therefore, SRT posits that the macrolevel division of labor leads to microlevel processes that result in different gendered behaviors (Diekman and Schneider 2010). Thus, women came to be thought of as more communal due to their historically higher levels of domestic work and childcare which then translated into more feminine, communal occupations which reinforces stereotypes of women being more communally inclined (Eagly et al. 2000; Koenig and Eagly

2014; Wood and Eagly 2002). Following the same logic, men are therefore thought of as more agentic because of their historically higher presence in the paid labor force which translates to men being more represented in masculine, agentic occupations which reinforces stereotypes of men being more inclined to agentic behaviors (Eagly et al. 2000; Koenig and Eagly 2014; Wood and Eagly 2002). However, how people enact gender roles can vary greatly based on the situation and the individual (Wood and Eagly 2002). Wood and Eagly (2002) posit a biosocial approach in which the most consistencies in the gendered division of roles and tasks occurs in the roles and tasks that are most closely associated with biological processes such as reproduction and that there are less and less gendered consistencies in roles and tasks that involve less biological processes.

Furthermore, gendered roles create contrasting expectations of behavior for men and women that are separate from any inborn differences between the sexes but that become institutionalized and are then reinforced in societal structures (Eagly et al. 2000). Johnson et al. (2008) employed both surveys and an experiment to determine the extent to which sex role expectations result in different expectations for male and female leaders with particular emphasis on demeanor and emotion displays. Through their mixed-methods approach, they find that people do have different role prototypes for male and female leaders with male leaders being more likely to be associated with agentic traits and female leaders more likely to be associated with communal traits (Johnson et al. 2008).

Fox and Oxley (2003) utilize data from state-level elections in the United States to examine how gender stereotyping may influence women's likelihood of running for and winning state executive office elections. The history of men working in public extra-familial spheres and women working in the private intra-familial sphere both creates and reinforces gender

stereotypes (Fox and Oxley 2003). The gendered division of labor leads to stereotypes about the characteristics of men and women which in turn are attributed to male and female political candidates (Fox and Oxley 2003). They find that while the gender of a candidate does not accurately predict who will win elections, women are much less likely to be nominated to run for positions that have more masculine role connotations such as comptroller, treasurer, and governor (Fox and Oxley 2003). In this way, gender role expectations are based on a gendered division of labor, affecting who makes it onto political tickets and therefore who ends up in leadership positions.

Role Congruity Theory

According to SRT, men are typically associated with prescriptive norms such as agency, assertiveness, and dominance while women are typically associated with prescriptive norms such as communality, deference, and obedience (Wood and Eagly 2002). However, that does not mean that people cannot occupy roles that are traditionally inconsistent with their perceived gender. Eagly and Karau's (2002) Role Congruity Theory of prejudice toward female leaders proposes that the perceived incongruity between the feminine gender role and leadership roles leads to sanctions and/or negative feedback from others. Role Congruity Theory (RCT) considers the congruity between gender roles and other roles, particularly leadership roles, as well as specifying key factors and processes that influence congruity perceptions and their consequences for prejudice (Eagly and Karau 2002).

RCT posits that when a person occupies a specific role that is incongruent to their gender role, they will likely receive sanctions and/or negative feedback from others due to the perceived inconsistency (Diekman and Eagly 2008). Therefore, when women act in agentic ways, they may be negatively evaluated and receive sanctions due to their perceived role incongruity. When a

person expresses traits that are not associated with their gender presentation such as a man-presenting person acting communally or a woman-presenting person acting agentic or when a person occupies a role that is not associated with their gender presentation such as a man who is a nurse or a woman who is a CEO, the evaluations of others may be negatively affected because of the incongruity (Eagly and Karau 2002; Eagly et al. 2000; Heilman 2012). Thus, women are seen as being acceptable leaders when the leader roles are more communal in nature such as dealing with children and family problems, helping the poor, and/or working for peace (Eagly and Karau 2002). However, when women are effective leaders in non-communal contexts they violate feminine gender norms and invoke masculine, agentic qualities which may lead to unfavorable evaluations or other forms of backlash due to the violation of their gender roles through their occupation of masculine leadership roles (Eagly and Karau 2002).

Garcia-Retamero and López-Zafra (2006) found that in workplace contexts, the gender role congruity of a leadership candidate affects whether or not others believe that the person will be a successful leader and attributions of success and failure also vary by the gender role (in)congruity of a candidate. Their results from show that participants predicted that males would be more successful in obtaining leadership positions in male (auto manufacturing) and unspecified (general manufacturing) industries and females would be more successful in obtaining a promotion to a leadership role in a female industry (clothing manufacturing). Additionally, participants tended to attribute the success of female candidates in incongruent or unspecified industries to external causes and female success in congruent industries and all instances of male success to internal reasons (Garcia-Retamero and López-Zafra 2006). Participants also attributed all failures to external causes except women working in incongruent industries, in this case they tended to attribute the failure to something internal to the woman

(Garcia-Retamero and López-Zafra 2006). These results indicate that not only do people tend to think that an individual will be a more successful leader in industries when the industry is congruent to their gender role and that people tend to consider the role incongruity of the target when attributing reasons for success and failure. Women working in gender role incongruent industries were more likely to have failures attributed to an internal cause than men or women in gender role congruent industries who were more likely to have failures attributed to external causes. Thus, role incongruent women receive more personal negative feedback when there is a failure than their role congruent women and men peers.

Status Incongruity Hypothesis

Rudman et al. (2011) argue that there are three main limits to RCT: (1) RCT does not account for backlash received by agentic women who are not in leadership roles and that it does not account for the risk atypical men experience; (2) RCT only broadly defines gender roles without specifying which aspects of gender roles are at fault in backlash; and (3) RCT does not specify the perceivers' motives for penalizing the target even though evidence suggests that motives are required for a person to engage in backlash, strong negative reactions that may be exhibited through criticism, disgust, and/or other negative responses with the to maintain the status quo of the traditional gender hierarchy (Brescoll, Okimoto, and Vial 2018; Rudman et al. 2011). They propose the Status Incongruity Hypothesis to mitigate these gaps in RCT.

Status Incongruity Hypothesis (SIH) is similar to RCT in that they both posit that negative evaluations or reactions can occur when a person behaves in gender counter-stereotypical ways. According to SIH, much like SRT, there are different types of gender norms—prescriptive gender norms that dictate what men and women ought to be and proscriptive norms that dictate what men and women ought not to be (Moss-Racusin, Phelan, and

Rudman 2010). Backlash reactions occur when a person deviates from their prescribed gender norms and include both tangible responses, such as negative evaluations as described by RCT, as well as strong negative emotional responses to perceived gender role incongruity due to the perceived threat the incongruent person presents to the gender hierarchy. Backlash responses to gender role incongruity might include people viewing gender role incongruent individuals as less psychologically healthy and as being cold and hostile (Heilman 2012). Backlash may also be exhibited in the workplace through paying gender role incongruent people less than role congruent persons, hiring less gender role incongruent persons, and promoting gender role incongruent persons less (Heilman 2012). For example, men may be penalized when they are passive, disclose emotions, and/or have success in feminine domains as they would be viewed as acting in a gender role incongruent manner thus eliciting backlash through negative evaluations and sanctions from others (Moss-Racusin et al. 2010). Gender counter-stereotypical behaviors of men, such as behaving modestly, have otherwise been understudied as much of the focus of backlash effects research has been on women.

SIH posits that defending the gender hierarchy motivates backlash reactions towards people who violate gender norms (Moss-Racusin et al. 2010). Therefore, people who violate stereotypes that most strongly justify the gender hierarchy are most at risk of receiving backlash while violations of gender norms that are less related to a justification of the gender hierarchy may be less likely to receive backlash (Moss-Racusin et al. 2010). Under the traditional gender hierarchy, women are subordinate and less powerful than men. Agentic women face backlash because they challenge the legitimacy of the gender hierarchy and are criticized not for defying the feminine gender role but because they are violating their place in the gender hierarchy (Brescoll et al. 2018). For example, research has found that women experience backlash when

they violate gender stereotypes and try to be leaders or enter other masculine domains which threatens the gender hierarchy, thus leading to responses of moral outrage and backlash reactions such as being critiqued more than their men peers (Brescoll et al. 2018; Moss-Racusin et al. 2010).

Empirical research tends to find support for SIH. For instance, Rudman, Moss-Racusin, Phelan, and Nauts (2011) found support for SIH in a multi-part study. The results of their studies support the SIH proposition that women can be leaders without receiving backlash but also find that when women are agentic and therefore a threat to the gender hierarchy, they tend to be penalized and receive backlash. In their first study, they found that women are proscribed from masculine displays of dominance and/or high status and that this proscription is used as a justification for prejudice against agentic women (Rudman et al. 2011). Furthermore, they find support for the idea that backlash effects are not universally attributed to gender counter-stereotypical behaviors but instead depend upon if the counter-stereotypical behavior is seen as a threat to the gender hierarchy (Rudman et al. 2011). In their second study, they conducted an experiment in which job candidates were varied on their gender and whether they were agentic or communal. They found that highly competent and accomplished women candidates who displayed communality were rated as similarly likable and hirable to highly competent accomplished men candidates who displayed communality (Rudman et al. 2011). On the other hand, highly competent agentic women were evaluated as significantly less likable and less hirable than their agentic men counterparts (Rudman et al. 2011). The results of this study indicate that women can be competent and accomplished and be seen as just as hirable as the men counterparts, so long as they are not agentic and thus threatening the gender hierarchy. Their third and fourth studies find further support for SIH, finding that people who more strongly

endorse the gender hierarchy were more likely to penalize agentic women as compared to agentic men and that when participants were primed with a threat to their system, in this case their country[2], they were even more likely to penalize agentic women who were seen as a further threat to the declining system (Rudman et al. 2011). Finally, their fifth study found that agentic women leaders were more likely to be sabotaged by subordinates while agentic men leaders and low agency men and women leaders were less likely to be sabotaged (Rudman et al. 2011).

More recently, Brescoll, Okimoto, and Vial (2018) employed SIH to test if emotions and moral judgments affect how people react to gender counter-stereotypes. Through an experiment, they find that people were statistically significantly less likely to vote for a highly voluble woman candidate, meaning they dominate the discourse through incessant talking, than they were to vote for a highly voluble man or a woman with average volubility (Brescoll et al. 2018). The highly voluble woman was also met with more expressed moral outrage (e.g., contempt, disgust, revulsion, and disdain), indicating that moral outrage accounts for at least some of the reasons why people are less likely to vote for this candidate (Brescoll et al. 2018). Further analyses showed that the direct effect for volubility became non-significant when controlling for moral outrage while this was not the case for men candidates, thus adding support to their hypothesis that moral outrage is a large factor in why highly voluble women candidates are rated as less likely to be voted for than men and women with average volubility (Brescoll et al. 2018).

Limits of RCT and SIH

Taken together, researchers have found support for Status Incongruity Hypothesis and they contend that it is distinct from RCT and even that it fills in some gaps of RCT (Brescoll et

[2] The authors point to previous empirical research (Kay et al. 2009) that shows that there is a direct relationship between system-justifying actions and threats to one's country (Rudman et al. 2011). Thus, they utilize news articles that are positive and negative to manipulate the level of threat to one's country in order to examine how system level threats may lead to greater levels of system justification and thus harsher responses to those who are further threats to the system (Rudman et al. 2011).

al. 2018; Rudman et al. 2011). However, even with some of the gaps of RCT filled by SIH such as the inclusion of men and women outside of leadership positions, there are still limitations to both theoretical frameworks that need to be addressed. While RCT and SIH have been leveraged to study the interaction of gender and many other statuses such as a person's leadership status, occupational status, and politician status, there are several limits to these theories. One such limit is that RCT and SIH have been used almost exclusively to examine the interaction between specific statuses like occupation and the diffuse status of gender. However, people occupy many diffuse statuses other than gender such as race and age that are influential in most social situations to varying degrees, much like gender (Koenig and Eagly 2014). Therefore, there is reason to believe that other salient social statuses such as race, age, and class operate in similar ways to gender (Ridgeway and Correll 2004).

Gender is a highly salient social category that "involves cultural beliefs and distributions of resources at the macro level, patterns of behavior and organizational practices at the interactional level, and selves and statuses at the individual level" (Ridgeway and Correll 2004: 510-511; Ridgeway and Smith-Lovin 1999). Additionally, much like there is a historical pattern in the United States of gendered occupations that led to certain characteristics being associated with men and other characteristics being associated with women, occupations have also been historically divided by race and other statuses such as class and age. Concerning race, White persons and men, in particular, have traditionally been associated with more prestigious careers than women and Black or Hispanic persons. Furthermore, White persons, and men, in particular, have in the United States, historically held the majority of the most prestigious and powerful positions in our society from government leaders to CEOs and company owners. Therefore, after generations of a racial division of labor that is very similar to the gendered division of labor and

White dominance that is very similar to patriarchal-male dominance, it is reasonable to think that the traits and characteristics associated with the careers and social positions of White persons versus Black, Hispanic, and other raced persons came to be associated with these different racial groups in general. Thus, whether applicable to individuals or not, there are likely status stereotypes that have come to be associated with White persons that may be different for Black persons and that may also be different for Hispanic persons which may yet be different for persons of other races due to the historical occupational and status positions traditionally held by different raced and gendered groups.

Intersectional approaches can help to further interrogate how current patterns of gendered and raced statuses that have been created through a long history of gendered and raced occupations and statuses are exhibited through current behavior, belief, and value expectations. The concept intersectionality was introduced in the 1980s and calls for an examination of the dynamics of difference and sameness in the consideration of gender, race, and other axes of power (Cho, Crenshaw, and McCall 2013). Social statuses do not exist independently of one another and intersectional approaches call for richer analyses that examine multiple statuses and therefore more closely resemble "real world" circumstances. Many salient social characteristics such as gender and race exist concurrently with one another and almost constantly, meaning that they impact almost all social interactions. That is why it is particularly important to consider the effects of statuses such as race and gender both independently and in conjunction with one another (Ridgeway and Kricheli-Katz 2013). A person's race status can affect how others perceive them and their gender status can also affect how others perceive them and a person's combined race and gender status may lead to entirely different perceptions and expectations. Intersectionality refers to the idea that "the critical insights that race, class, gender, sexuality,

ethnicity, nation, ability, and age operate not as unitary, mutually exclusive entities, but as reciprocally constructing phenomena that in turn shape complex social inequalities" (Collins 2015: 2).

Intersectional approaches can help to illuminate situations in which groups that are generally assumed to be homogenous, such as women or Black persons, actually have great variability within them. There is empirical evidence that suggests that people hold different behavioral expectations for men and women of different races. Livingston, Rosette, and Washington (2012), for example, conducted an experiment and found that an agentic Black woman leader was evaluated more positively than an agentic White woman leader and an agentic Black man leader. In fact, the agentic Black woman leader was evaluated most similarly to agentic White men while the agentic Black man leader was evaluated most similarly to the agentic White woman leader (Livingston et al. 2012). These results indicate that role congruity gender expectations may be mitigated by the race of the person in question. Non-intersectional approaches presume that all women and all men are generally perceived and evaluated the same way by others and therefore would have missed the complex ways in which race and gender combine to affect perceptions and evaluations that Livingston et al. found (2012).

Clearly, the complex ways in which race, gender, and other social statuses combine to affect perceptions and evaluations needs to be examined more closely as the effects can be profound and yet intersectional analyses are often lacking in social psychological research (Hollander and Howard 2000; Hunt et al. 2013). Role Congruity Theory and Status Incongruity Hypothesis both currently suffer from this very problem. The lack of integration of salient social statuses other than gender such as race and the lack of intersectional research approaches may mean that previous RCT and SIH research has not fully examined the complex ways in which

diffuse statuses and specific statuses interact to affect perceptions and expectations. Through the incorporation of more social statuses and the use of intersectional frameworks, RCT and ideas from SIH can be leveraged to examine more complex status dynamics that more closely reflect real world circumstances.

Study Context: Student Evaluations of Instruction

RCT and SIH both provide explanations as to why a person who is in a role that is incongruent with their gender may receive sanctions such as negative evaluations from others. With the addition of intersectionality, RCT and SIH can be extremely versatile theoretical frameworks for the examination of how multiple statuses affect the content and completion of subjective evaluations. Student evaluations of instruction (SEIs) are one specific example of subjective evaluations that are used to evaluate individuals who may be affected by role (in)congruity.

SEIs were introduced to higher education in the United States in the 1920s and have since become a nearly ubiquitous part of United States higher education (Algozzine et al. 2004; Benton and Cashin 2014). SEIs provide a relatively simple means for institutions of higher education to collect data on the overall course quality and the effectiveness of the instructors directly from the students in a course (Benton and Cashin 2014). However, SEIs can also be problematic if they are not tested for reliability and validity, if they are administered inconsistently within and between departments, and if they are the only source of data regarding teaching effectiveness used in instructor evaluations (Benton and Cashin 2014). Even with these problems, SEIs are commonly used in the retention, tenure, and promotion process and can carry more weight than other factors that are considered (Clayson 2009; Franklin 2001).

Status-Based Biases in Student Evaluation Results

In fall 2019 the American Sociological Association put out a "Statement on Student Evaluations of Teaching" in which they outline the severe gender and racial discrepancies that have been shown to persist in student evaluations of teaching as well as steps that could be taken to address these systemic disparities that advantage White and men faculty over women and faculty of color (Anon 2019). As of July 2020, the statement from ASA has been endorsed by twenty-two other professional organizations including the American History Association, the Canadian Sociological Association, the American Political Science Association, and the National Communication Association (Anon 2019). The widespread support from other professional organizations of the ASA's statement indicates not only the continued disparities in student evaluations of teaching across a myriad of disciplines but also the desire to address these disparities through systemic change across disciplines.

Research consistently suggests that there are gender and race disparities in student evaluations of instruction (SEIs) (Basow 1995; Bavishi, Madera, and Hebl 2010). One area that has been studied quite extensively is gender biases in SEIs (Basow 1995; El-Alayli, Hansen-Brown, and Ceynar 2018; MacNell, Driscoll, and Hunt 2015). The results of many studies show that women instructors tend to receive lower evaluations than men instructors (Basow 1995; Boring et al. 2016; El-Alayli, Hansen-Brown, and Ceynar 2018; MacNell, Driscoll, and Hunt 2015a). Some of these gender differences can even lead to more effective women instructors being evaluated worse than their less effective men peers (Boring et al. 2016).

Expanding upon the idea that teacher quality may not matter as much as the statuses of the instructors, MacNell, Driscoll, and Hunt (2015) utilized a two-by-two between-subjects experimental design in which two teaching assistants (one male, one female) taught two sections of the same online course where the instructor retained their sex/gender identity in one section

but assumed the sex/gender identity of the other instructor in the other section. Their design allowed the researchers to hold teaching style constant so that the perceived gender of the instructor could be isolated in their analyses. Students rated the perceived male instructor higher than the perceived female instructor, regardless of the actual gender of the instructor (MacNell et al. 2015). Their results indicate that there are gender biases in student evaluations of instruction as the same actions by the instructors were perceived differently depending on the perceived gender of the instructor. Perceived female instructors were rated as less prompt (3.55 out of 5) than perceived male instructors (4.35 out of 5) even though grades were always posted at the same time across sections (MacNell et al. 2015). Additionally, male instructors were rated higher on all six interpersonal measures, indicating that female instructors may be expected to be more interpersonal whereas men are not and are therefore rewarded as having gone "above and beyond" when they do display interpersonal traits (MacNell et al. 2015). The authors conclude that "the combination of higher expectations and lower automatic credibility translates into very real differences in student ratings of female versus male instructors" (MacNell et al. 2015: 300). These results indicate that even when holding teaching style, grading, and course matter consistent, students' evaluations of instructors are still affected by faculty gender.

Less research has examined the effects of faculty race and other statuses on SEIs as compared to gender, however, the research that has been done suggests that faculty who are persons of color may be disadvantaged on SEIs as compared to their White counterparts much like women are disadvantaged as compared to men. Reid (2010) examined student evaluations of teaching from RateMyProfessors.com and found that the best-ranked instructors were more likely to be White while the worst-ranked instructors were more likely to be Black or Asian. Smith and Hawkins (2011) evaluated multiple years of SEI data to compare student evaluations

of faculty in three racial groups: White, Black, and Other. They found that of the three racial groups, Black faculty mean evaluation scores were the lowest of all of the groups across 28 items (Smith and Hawkins 2011). Anderson and Smith (2005) conducted an experiment in which students rated a hypothetical course and instructor based on course syllabi that varied by teaching style, professor gender, and professor ethnicity. They found several interaction effects on students' evaluations of faculty based on the teaching style, gender, and ethnicity of the faculty member indicating that different course and faculty characteristics lead to differences in students' subjective evaluations of instructors (Anderson and Smith 2005). For example, they found that White women with a strict teaching style were viewed as warmer than Latina professors with a strict teaching style while Latina professors with a lenient teaching style were viewed as more warm than White women with a lenient teaching style (Anderson and Smith 2005). The results of their study illustrate the importance of considering how instructors' multiple intersecting statuses may affect how students perceive and evaluate them.

Bavishi, Madera, and Hebl (2010) and Basow (1995) both take SEI research a step further by including the discipline being taught in their studies. Bavishi, Madera, and Hebl (2010) find that based on hypothetical CVs, students perceive faculty of different races and faculty in different disciplines differently. Specifically, White instructors tended to be ranked higher in competence and legitimacy than Black or Asian instructors and science instructors tended to be ranked as more competent and legitimate than humanities instructors (Bavishi et al. 2010). While they did not find any gender differences, their results do indicate that not only can faculty statuses like race affect SEIs but so can the discipline in which a faculty member is teaching.

Basow (1995) examined SEI data across two years and found a significant three-way interaction between teacher gender, student gender, and discipline for fourteen questions about teaching behaviors as well as significant two-way interactions between teacher gender and student gender and between teacher gender and discipline. Men teachers received statistically significantly higher ratings than female teachers for all questions except for sensitivity and student comfort, both of which are more communal, feminine qualities (Basow 1995). Female students' ratings were statistically significantly higher than male students' ratings and humanities teachers received the highest ratings, while natural science teachers received the lowest ratings on all questions (Basow 1995). The significant interaction between teacher gender and discipline indicates that in the humanities women are rated similarly or higher (enthusiasm, student freedom, non-repetition, and feedback) than their male colleagues on all questions, in the natural sciences women are rated slightly lower than men on all questions, and there are mixed results in the social sciences with men rated slightly higher on some measures (overall, appropriate speech, enthusiasm, thought stimulation, organization, non-repetition, and knowledge) and women professors rated higher on other measures (sensitivity, respect, fairness, and student freedom) (Basow 1995). The course questions generally showed that men instructors were rated more positively than women instructors and this was strongest in the natural sciences (Basow 1995). Thus, their results show that in disciplines that are traditionally more masculine, women instructors tend to be rated lower than their man counterparts whereas in disciplines that are traditionally less masculine, women instructors are rated as positively or better than their man counterparts. This indicates that gender role congruity expectations of the instructor of a course may affect students' subjective evaluations of that instructor, particularly when the instructor is perceived to be gender role incongruent with the discipline they are teaching.

Significance of the SEI Context

Occurring concurrently with research that continues to find status-based differences in SEI scores and differences in the statuses of SEI completers, initiatives such as the National Science Foundation's ADVANCE and AGEP grant programs have been working to increase the diversity of faculty statuses in higher education. These programs include increasing the number of women and faculty of color in STEM fields (science, technology, engineering, and mathematics) as STEM fields have been particularly dominated by White men (Blackburn 2017). The success of ADVANCE, AGEP, and other programs aimed at increasing faculty diversity means that there are now more women faculty and faculty of color than ever before working in disciplines that are not traditionally associated with their gender, race, or other social roles (Davis and Fry 2019). And yet, even with these equity programs, there is persistent evidence that suggests that instructors who violate role congruity expectations may be evaluated more harshly by their students than instructors who are perceived to teach in role congruent disciplines (Basow 1995; Bavishi, Madera, and Hebl 2010). Programs such as AGEP and ADVANCE are purposefully working to increase the representation of women and faculty of color in higher education with a particular emphasis on STEM fields, but if traditional SEIs are biased against women and/or faculty of color in general, they may be even more biased in STEM fields where there is a particular dearth of women and faculty of color.

If traditional SEIs tend to produce results that are biased against minority status groups in higher education such as women and faculty of color, as has been found by previous research (Basow 1995; El-Alayli et al. 2018; MacNell, Driscoll, and Hunt 2015b), and even more so in STEM fields, the preexisting gender and racial inequalities in higher education may be perpetuated and accentuated through the use of traditional SEI data in the retention, promotion, and tenure processes (Clayson 2009; Franklin 2001). Thus, even if programs such as AGEP and

ADVANCE are successful in onboarding more women and faculty of color into faculty positions in STEM disciplines, these faculty may leave higher education due to biases in SEIs that affect their likelihood of being retained, promoted, and tenured. Therefore, it is crucial to understand how students may evaluate faculty of a variety of social statuses in a variety of academic disciplines similarly or differently from one another and how these evaluations may work against diverse instructors in higher education.

CURRENT RESEARCH

The role of professor has been historically gendered and raced in the United States such that White men have predominately filled this occupational role. However some disciplines, such as STEM fields, have been and continue to be more dominated by White men than by women and persons of color (Blackburn 2017). Therefore, when a woman is a professor in STEM she not only has entered a more traditionally masculine occupation as a professor but is also in a more traditionally masculine field. Both levels of gender role incongruity may lead her students to see her as poorly fitting into the role of STEM professor which may in turn affect their expectations for her performance in the role and therefore, their evaluations of her in the role. As such, women who teach in STEM may receive more negative evaluations than their men counterparts or women who teach in fields more traditionally associated with women such as English. Men, on the other hand, may receive backlash when they teach in more woman-dominated fields such as English as compared to when they teach in more man-dominated fields such as the STEM disciplines.

As mentioned, little attention has been given to race or intersectionality by congruity scholars up to this point, but because race is also a salient social role that people automatically process upon meeting someone, it may operate in similar ways to gender (Ridgeway and Correll

2004). There is also evidence that suggests that people evaluate leaders differently if the leader is a White woman versus a Black woman versus a White man versus a Black man (Livingston et al. 2012). Therefore, it is crucial to take the intersectional identities of the targets of subjective evaluations into account when looking at between- and within-group differences. Black women may be more able to defy gender role expectations than their White women counterparts and therefore may not receive backlash effects while Black men may be penalized for entering into occupations that do not align with their racial status (Livingston et al. 2012). Therefore, students may evaluate their Black women faculty differently from their White women faculty and there may be even more differences depending on the discipline of the instructor in question. Thus, though the incongruities literature has not, to my knowledge, been used to examine the interactional effects between salient social statuses beyond gender and diffuse statuses, the incorporation of other salient social statuses constitutes an important contribution to the RCT and SIH literatures.

My dissertation applies RCT and SIH literatures to examine how students' subjective evaluations of their instructors are affected by the role (in)congruity of their instructors. For instance, is a woman teaching in engineering evaluated differently than a woman teaching in English due to the masculinity associated with engineering versus the femininity associated with English? Is there a similar effect on the evaluations of men who are instructors in traditionally feminine versus traditionally masculine fields? Does race/ethnicity interact with gender and discipline to affect students' perceptions and evaluations of their instructors? The first and second studies of this dissertation examine these questions through analyses of quantitative SEI data utilizing exploratory factor analyses (EFA), confirmatory factor analyses (CFA), multiple indicators multiple causes (MIMIC) models, and structural equation models (SEM). In the third

study, open-ended student evaluation questions are qualitatively examined to tease out potential causal mechanisms for any observed differences in student evaluations based on instructor gender, gender role (in)congruity, and/or race/ethnicity. The analysis both quantitative and qualitative SEI data allows for a more detailed study of how intersectional inequalities may affect student evaluations than what has been done in previous studies. In the following paragraphs I briefly summarize the main research question and the analyses completed in each of the three studies conducted. All three studies utilize the same student evaluation of instruction dataset which is described in more detail below.

Study 1 (Chapter 2): Quantitative Analyses of Student evaluations of instruction with Attention to Faculty Gender and Role (In)Congruity

In the first study, quantitative student evaluation data are used to answer the question: are students' subjective evaluations of their instructors affected by the gender and perceived gender role (in)congruity of their instructors? The analyses include exploratory factor analyses (EFA), confirmatory factor analysis (CFA), multiple indicators multiple causes (MIMIC) models, and structural equation models (SEM) which were conducted on all available student evaluation data from five fall and spring semesters at a large land-grant and research-intensive university in Appalachia.

Study 2 (Chapter 3): Quantitative Analyses of Student evaluations of instruction with Attention to Faculty Gender, Role (In)Congruity, *and* Race

In the second study, the quantitative analyses from Study 1 are expanded upon to answer the question: are students' subjective evaluations of their instructors affected by the gender, perceived gender role (in)congruity, and/or race/ethnicity of their instructors? Grouped structural equation models on the same data set used in Study 1 (Chapter 2) were used to answer this question.

Study 3 (Chapter 4): Qualitative Analyses of Open-Response Questions from Student evaluations of instruction to Investigate Potential Causal Mechanisms of Gender and Race Differences

Finally, in Study 3 qualitative analyses of open-ended SEI questions are analyzed to further answer the question: are students' open-ended subjective evaluations of their instructors affected by the gender, perceived gender role (in)congruity, and/or race/ethnicity of their instructors? Coding was completed on 1,430 open-ended student evaluation responses across six themes: positive personal, negative personal, positive professional, negative professional, positive course, and negative course. The codes were then quantitatively analyzed to determine which instructor groups receive the most of each type and the most comments overall. The responses were also qualitatively analyzed using sentiment analysis, word clouds, and code summaries.

DATA

To examine the effects of perceived instructor gender role congruity on SEIs, Student evaluation of instructors (SEI) data from eleven semesters were obtained from a large research-intensive institution in the Appalachian region of the United States. Data from the summer and winter terms were eliminated from the sample as there is reason to believe that summer and winter classes, which are considered highly optional and considerably shorter, may operate differently than traditional fall and spring courses. The spring 2016 data was also removed as the institution moved to exclusively online evaluations and this semester occurred before the move to all online SEIs thus, there may be substantial differences in data collection between this term which was collected in person and the other terms which were collected online. Therefore the dataset includes five semesters of fall and spring term data which was exclusively collected online from the terms from fall 2016 to fall 2018.

The data includes information about student, course, and instructor characteristics. The SEI data includes student information such as their gender, if they are an international student, which college they are a part of, their class standing, if they are an athlete, if they are a first-time freshman, their class standing, and their GPA. The SEI data also includes information about the course such as the subject, the college the course is housed in, the course number, the course type (lecture, lab), the instructional method (web-based, in-person), the times and dates the course met, if the course satisfies a general education requirement, and if the course is restricted to only students with certain majors. Furthermore, there is information about the instructors of the course such as their department, their title, and if they were ever a student at the institution. The data were received in their raw output form directly from the institution, thus an extensive data cleaning and merging process occurred before beginning analyses.

Data Merging and Cleaning

The student evaluation data were obtained in five separate datasets, one per semester of data. The data were merged into one complete SEI dataset. To merge the datasets, all the column titles were compared to combine like columns.[3] The data cleaning and variable addition processes that follow in this section were completed on the combined data which includes the data from all five semesters. In addition to the SEI data, data about the characteristics of the instructors were obtained from the institution's human resources department. The data from

[3] In the process of merging the five datasets, it was discovered that there was a significant change in the student evaluation forms between the spring 2017 and fall 2017 semesters. Due to this change, only the latter three semesters (fall 2017 through spring 2018) are included in the quantitative analyses presented in Chapters 2 and 3. The earlier two semesters of data (fall 2016 and spring 2017) were not viable for the quantitative analyses due to a lack of consistency in the questions being asked which led to low sample sizes. However, the data cleaning and variable addition processes that are described in this section were completed on data for all the semesters. These processes were completed on all the data so that it was performed consistently on all the data as the earlier two semesters were utilized in the qualitative analyses described in Chapter 4. Completing all of the data cleaning now would also allow for follow-up quantitative analysis to occur.

human resources included instructor gender, race, home department, position title, and more. The human resources data and the complete SEI dataset were merged into one master dataset. The two datasets were attached through course reference numbers (CRNs) which are unique to each course and were present in each dataset. The statistical program R was used to automate the process of matching and merging the two datasets. The matching and merging process resulted in one dataset that could then be cleaned, and to which variables could be added prior to beginning analyses.

To begin the data cleaning process, the data was reduced to include only undergraduate courses. The next step was to remove certain instructors from the dataset. Graduate teaching/research/general assistant instructors were removed from the sample as there is reason to believe that students may respond differently to graduate student instructors than they would to other instructors. SEI responses for professional schools at the university were also eliminated from the sample. There is reason to believe the ways in which students and faculty interact in professional schools may be substantially different from other colleges. These professional schools include the schools of pharmacy, dentistry, medicine, nursing, and public health. Follow-up studies could include examining the SEI scores for graduate-level classes, graduate student instructors, and/or professional schools.

Variable Creation

Variables were then recoded and added to the merged and cleaned dataset as needed. Two variables which were recoded include the instructor race/ethnicity variable and instructor gender variable which were both originally from the human resources data. The way in which the race/ethnicity variable is coded in the original data was somewhat problematic. Although labeled as “race”, due to the categories listed it is actually conflating race *and* ethnicity in one variable.

The categories recorded were White, Black, Asian, Hispanic, American Indian/Alaska Native, Native Hawaiian/Pacific Islander, two or more races, and unknown. To have sufficient sample sizes across instructor genders and disciplines, the race/ethnicity variable was recoded into two groups: White and racially/ethnically minoritized. Instructors whose race was labeled as "unknown" were dropped from the sample as they could not be added to either the White or racially/ethnically minoritized category. Similarly, the variable for instructor gender was measured in a somewhat problematic way. The variable in the human resources data is labeled "gender", but the categories provided are actually sexes with the options being male, female, and unknown. Instructors whose gender was labeled as "unknown" were also dropped from the sample as they could not be added to either the male or female category. These variables clearly conflate gender and sex as well as race and ethnicity which are each distinct attributes. However, this is one of the limitations of utilizing secondary data and ultimately as these are self-reported, they should not have a substantial effect on the results of the study.

Data regarding the national gender distribution of faculty by discipline was not found. Thus, data from the 2020 National Science Foundation and National Center for Science and Engineering Statistics' Survey of Earned Doctorates (National Center for Science and Engineering Statistics 2021) was used to approximate the gender dominance of professors in each discipline. One of the available datasets from the survey includes the sex and major field of study of doctorate recipients for each year from 2010 to 2020. The 2020 data from this dataset was utilized to calculate the percent of male and percent of female doctoral recipients in each major field of study. Disciplines in which there was greater than 55 percent of one gender were marked as being gender dominant for professors in that discipline. Disciplines in which there

were between 45 and 55 percent of both genders were marked as being gender neutral[4]. Thus, three categories for gender dominance were determined based on this data: man-dominate, woman-dominate, and neutral.

In cases where there was not an exact match between the Survey of Earned Doctorates category and a major at the institution being studied, the closest major possible from the Survey of Earned Doctorates to that of the institution were used to determine the gender-dominance category for that discipline. The gender-dominance categorization of disciplines can be found in Appendix 1. These gender dominance labels were then added to the SEI data based on the home department of the instructor of the course. The doctorate earner data was used as an approximation for faculty gender distribution by discipline because the people earning doctorates in a given field are typically the possible candidates for faculty positions in that field. Thus, the gender and discipline distribution of doctorate earners in a given field should be indicative of the gender distribution of new faculty in said field. While this may not be wholly accurate as not all doctorate earners work in academia and it does not reflect the pre-existing gender domination of disciplines, it is reasonable to suspect that the gender distribution of the doctorate earners and faculty in each field are not egregiously different.

The variables for instructor gender, instructor department, and the gender dominance of the discipline were used to create a variable for instructor role congruity. The role congruity

[4] In this dissertation, disciplines in which there are between 45% and 55% of both men and women are referred to as "gender role neutral" because they constitute a fairly even split of both men and women doctoral recipients. However, these disciplines may also be considered "gender role balanced" as there is an about equal amount of both men and women experts thus making them balanced between the two groups. The choice to refer to these disciplines as gender role neutral affects the interpretations of the results such that they would be different if the disciplines were referred to as gender role balanced. This dichotomy of perspectives illustrates the subjectivity of the research process even within quantitative research and shows how even seemingly small decisions throughout the research process need to be carefully and thoroughly considered. Because this research was conducted with "gender neutral" in mind, the results that follow are presented through that lens, however, this is only one potential perspective through which these results could be examined.

variable has the categories role congruent, role incongruent, and role neutral. When an instructor's gender matched the gender-dominance of the discipline in which they teach as indicated by their listed home department, they were marked as role congruent. When an instructor's gender did not match the gender-dominance of the field in which they teach, they were marked as role incongruent. Finally, when an instructor taught in a gender-neutral discipline they were marked as gender neutral. For example, a woman teaching in English (woman-dominated) was marked as role congruent while a woman teaching in engineering (man-dominant) was marked as role incongruent and a woman teaching in marketing (neutral) was marked as gender-role neutral. Three dummy variables were created wherein each category of role congruity was set to be equal to one and all else was set to be missing. These dummy variables allowed for easy selection of one group of faculty congruity at a time during analyses. The merged, cleaned, and amended dataset described here was used to complete the analyses in all three of the studies that follow.

CHAPTER 2: QUANTITATIVE ANALYSES OF STUDENT EVALUATIONS OF INSTRUCTION WITH ATTENTION TO FACULTY GENDER AND ROLE (IN)CONGRUITY

INTRODUCTION

Perceptions of target role congruity has been shown to affect subjective evaluations in a variety of areas such as dating (Hitsch et al. 2010), employment potential (Foschi et al. 1994), and leadership ability (Garcia-Retamero and López-Zafra 2006; Smith et al. 2019). Researchers have found that being perceived as role incongruent tends to have negative effects on a person such as being viewed as less competent than role congruent peers which is communicated through sanctions and/or negative feedback (Diekman and Eagly 2008; Fassiotto et al. 2018). For example, Fassiotto et al. (2018) found that medical residents tended to rate their female physician faculty lower than their male faculty across all specializations but to an even greater extent in specializations that were particularly male-dominated. Their findings suggest that even highly competent women such as physicians may be penalized through lower subjective evaluations from trainees due to their perceived role incongruity. I extend the examination of the effects of perceived role congruity on subjective evaluations even further through the investigation of the student teaching evaluations in higher education which, to my knowledge, is a novel area of exploration.

Teaching evaluations are a type of subjective evaluation that have become almost ubiquitous in higher education in the United States wherein students evaluate their instructors at the end of each school term. Frequently called Student Evaluations of Instruction (SEIs), SEI scores have been shown to vary widely depending upon students' reactions to faculty characteristics such as gender and race (Boring et al. 2016; Smith and Hawkins 2011). Discrepancies in SEI scores between groups of faculty based on their statuses and not their teaching are incredibly problematic as these scores are frequently used during the retention,

tenure, and promotion processes. Considerably less work (Basow 1995) has examined if women faculty are doubly penalized when they teach in a field that is perceived as being incongruent with their gender. If biases exist in the subjective evaluations of individual instructors because of gender and discipline "fit," then entire groups of instructors are poised to receive negative evaluations, regardless of *actual* teaching quality, that could inhibit their retention, tenure, and promotion in higher education.

In this study, student evaluations of instruction are quantitatively analyzed with consideration of the gender and perceived gender role (in)congruity of the course instructor. Through these analyses, I seek to answer the question are students' subjective evaluations of their instructors affected by the perceived gender role (in)congruity of the instructor? In the next section, I outline one theory of congruity, Role Congruity Theory (RCT), as well as previous research on student evaluations of instruction (SEIs).

LITERATURE REVIEW

Role theory is a social psychological theory that considers the positions a person occupies, the expectations of those roles, and how those roles affect their own and others' behaviors, attitudes, and values (Jacobs 2018a). All people occupy multiple roles and they must learn the expectations of each individual role (Jacobs 2018a). Roles occur at different levels, from specific roles which occur in very particular situations to diffuse roles which occur in most situations (Diekman and Schneider 2010a). Diffuse roles include statuses such as gender, race, and age because they are influential across almost every social situation (Koenig and Eagly 2014). Occupation and parental status are specific roles because they occur in particular circumstances and are not necessarily influential in other situations (Diekman and Schneider 2010a).

The demands of a person's diffuse roles can affect the extent to which their behavior is determined by their specific roles and vice versa. A specific role can become a very influential portion of a person's identity and the more internalized they become, the more likely a person is to carry out the expectations associated with that role over the expectations of other roles (Diekman and Schneider 2010a). For example, stay-at-home parents may experience the specific role of parent very differently from parents who work full time because the parent role may be less internalized by a person who also carries an occupation as a specific role. Additionally, people may have different role expectations for men versus women (diffuse) who are in the same occupation (specific) such that even if they execute the same occupational task, it may be perceived differently due to the difference in their diffuse roles (Eagly et al. 2000).

Social Role Theory (SRT) is an extension of Role Theory which focuses on how a person is affected both internally and externally by the multiple roles they occupy, especially when those roles carry conflicting expectations (Eagly and Karau 2002). SRT posits that the varying distribution of men and women into different social roles explains gendered differences in behaviors and personalities (Eagly et al. 2000; Koenig and Eagly 2014). Social roles are the shared expectations of persons who occupy a certain social position or who are members of a particular social category (Eagly et al. 2000). When a particular group is overrepresented in a social role, perceivers come to believe that the behaviors of that group are then generalizable to everyone in the group thus creating a group stereotype (Koenig and Eagly 2014). Women, for example, have been overrepresented in childcare roles thus leading to the persistent stereotype that women are warm, communal, and nurturing (Koenig and Eagly 2014).

Role Congruity Theory

Role Congruity Theory (RCT) extends SRT further by examining what happens when a person occupies multiple roles that are "incongruent" with one another such as women who take on masculine occupational roles (Eagly and Karau 2002). RCT argues that when a person occupies a specific role with expectations that are incongruent with their diffuse gender role, they will receive sanctions and/or negative feedback from others (Diekman and Eagly 2008). For example, if a man-presenting person acts communally—warm, caring—or a woman-presenting person acts agentic—assertive, analytical—perceivers will tend to evaluate them negatively due to the perceived role incongruity (Eagly and Karau 2002; Eagly et al. 2000; Heilman 2012a).

Evaluations of role (in)congruity have been shown to affect perceptions and evaluations of men and women in a variety of areas such as leadership, politics, and work (Diekman and Schneider 2010a; Fox and Oxley 2003; Garcia-Retamero and López-Zafra 2006; Simpson 2004a). Much role congruity theory research focuses on the effects of women who occupy specific masculine roles with particular emphasis on leadership and work (Brescoll, Okimoto, and Vial 2018; El-Alayli et al. 2018; Fox and Oxley 2003; Heilman 2012a; Johnson et al. 2008; Rudman and Glick 2001a). Less research has examined the effects of role incongruity among men, but this is a growing area of research that has started to examine the effects of occupational role incongruity (Johnson et al. 2008; Simpson 2004a).

Study Context: Student Evaluations of Instruction

Student evaluations of instruction (SEIs) are one specific example of subjective evaluations that are used to evaluate individuals that may be affected by gender role (in)congruity. SEIs were introduced a century ago and have since become a nearly universal practice in higher education in the United States (Algozzine et al. 2004; Benton and Cashin

2014). SEIs ask students their opinions of courses to evaluate the teaching of instructors and the various factors that may affect said teaching (Algozzine et al. 2004). SEIs are frequently used as evidence for or against instructors in the hiring, retention, tenure, and promotion processes and often carry more weight than other factors that are considered (Clayson 2009; Franklin 2001).

However, SEIs have been found to be exceptionally problematic. The American Sociological Association put out a "Statement on Student Evaluations of Teaching" in 2019. In this statement, they outline the severe gender and racial discrepancies that occur in student evaluations of teaching and call for changes to be made to SEIs to rectify these problems (Anon 2019). This statement has gained widespread support from twenty-two other professional organizations thus indicating that the problems associated with SEIs are far-reaching within academia and so is the call for changes to be made to the student evaluation of instructors.

Multiple studies have found that there are gender differences in student evaluations of instruction such that women tend to be rated lower than their men colleagues (Basow 1995; Boring et al. 2016; El-Alayli et al. 2018; MacNell et al. 2015b). For example, Boring et al. (2016) found that not only are SEIs statistically significantly biased against female instructors, but these gender biases can be large enough to cause more effective instructors to receive lower SEI scores than less effective instructors. Their results indicate that more effective instructors may receive lower SEI scores than less effective instructors simply because of gender-based biases (Boring et al. 2016). They use Centra and Gaubatz (2000:17) to define bias as occurring when "a teacher or course characteristic affects teacher evaluations, either positively or negatively, but is unrelated to criteria of good teaching such as increased student learning". Therefore, their results indicate that women instructors tend to receive more negative student evaluations because of their gender identity and not because students learn less in their courses.

MacNell, Driscoll, and Hunt (2015) conducted a two-by-two experiment in which they varied the actual and presented gender of the instructor of online courses to test this assertion. Two instructors, one male and one female, each taught two online sections of the same course, one presenting as a man and one presenting as a woman. Through this experiment, they found that even when all else including grading procedures, communication, and teaching effectiveness/style is held equal, the gender presentation of the instructor affects how students evaluate them on SEI forms (MacNell et al. 2015b). Students rated the perceived male instructors higher than the perceived female instructors, regardless of the actual gender of the instructor (MacNell et al. 2015b).

Even the same actions can be perceived differently depending on the perceived gender of the instructor. Returning to MacNell et al. (2015b), perceived female instructors were rated as less prompt (3.55 out of 5) than perceived male instructors (4.35 out of 5) even though grades were always posted at the same time across the four sections. The authors conclude that "the combination of higher expectations and lower automatic credibility translates into very real differences in student ratings of female versus male instructors" (MacNell et al. 2015; 300). Their findings, taken together with other research on gender biases in SEIs, indicate that there are definitely gender-based biases occurring in how students evaluate their instructors on student evaluation forms.

While much research indicates that there are gender disparities in SEIs, significantly less research examines how these gendered effects may vary by the discipline of the instructor being evaluated. Basow (1995) found statistically significant interactions between instructor gender and discipline. specifically, men instructors tended to receive statistically significantly higher ratings than women instructors on almost all SEI questions (Basow 1995). But, the results were

moderated by discipline such that women instructors in the humanities tended to be rated similarly to or higher than men in the humanities on all SEI questions whereas women in the natural sciences were rated slightly lower than men in the natural sciences on all questions (Basow 1995). The results for the social sciences, which tend to be more gender-neutral, were mixed such that men scored higher on some SEI measures and women scored higher on others (Basow 1995).

Basow and Montgomery (2005) also find that student evaluations vary by instructor gender and discipline. Female professors were rated significantly higher than male professors in the humanities and natural sciences but lower than male professors in the social sciences (Basow and Montgomery 2005). However, they find that in general professors in natural sciences score lower than other disciplines. Female professors in the natural sciences, though they score higher than male professors in the natural sciences, score significantly lower than female professors in the humanities but not social sciences (Basow and Montgomery 2005). These mixed results indicate that differences in student evaluation scores are more complicated than just varying by gender or discipline but rather both must be examined in conjunction with one another. The results of both of these studies highlight the importance of not examining differences in student evaluation scores with faculty characteristics in isolation but rather the need to consider how student perceptions may be affected by the ways in which instructor identities intersect.

These results indicate that in the current study, gender role (in)congruity of the instructor with respect to the discipline they teach in may affect the SEI scores received. Faculty are considered gender role congruent when they teach in a discipline in which their gender is in the numeric majority. Therefore, women who teach in humanities, education, and other women-dominated disciplines are considered to be role congruent whereas men teaching in those

disciplines are considered to be role incongruent. Men teaching in STEM (science, technology, engineering, and math) and other man-dominated disciplines are considered to be role congruent whereas women teaching in those disciplines are considered to be role incongruent.

Women teaching in STEM fields and men teaching in humanities may be perceived to have a "lack of fit" between their gender and their career (Heilman 2012a). Women and men who are role incongruent are likely to be penalized through negative evaluations (Eagly and Karau 2002; Heilman 2012a). Thus, according to role congruity theory, faculty who are perceived to be role incongruent, "lack fit", by their students may be penalized for their role incongruity in the form of lower SEI scores (Eagly et al. 2000; Heilman 2012a; Rudman et al. 2011). Role incongruent women may be even more likely to receive negative evaluations because they not only violate gender roles by teaching in man-dominated disciplines but they also defy gender roles by acting as a leader through being the leader of the classroom (Brescoll et al. 2018; Eagly and Karau 2002; Heilman 2012a; Johnson et al. 2008). Due to this double gender role violation, women instructors in role incongruent disciplines are likely to receive lower scores than men in role incongruent disciplines as these men are only violating one gender role and are not seen as violating leadership roles.

In this study, quantitative SEI scores are analyzed with consideration of the gender, discipline, and perceived role (in)congruity of the instructor to determine if perceptions of instructor gender role congruity affect how students evaluate their instructors. The SEI scores of women and men in women-dominated, men-dominated, and gender-neutral disciplines will be examined in this study. Therefore, this study not only focuses on more than just women as what tends to occur in both the congruity and SEI literatures, but it also considers that not all roles are masculine or feminine through the inclusion of gender-neutral disciplines. These two additions

constitute major contributions to the congruity and SEI literatures as they have, to my knowledge, been lacking thus far.

Furthermore, in addition to considering how both the gender and discipline of instructors might affect student evaluations, this study takes the analyses of SEIs a step further with respect to the statistical analyses conducted. In this study, the data are tested using multiple-indicators-multiple-causes (MIMIC) models and grouped structural equation models (SEMs) to determine if there is measurement error based on the gender and role congruity of the instructor. Measurement error refers to biases in how scales measure constructs depending upon construct-irrelevant group differences. In this study, SEI data is tested to determine if there is measurement error in the SEI forms based on the gender role congruity of the faculty being evaluated. Comparing the mean SEI scores of men and women instructors without determining if there is measurement bias may lead to incorrect conclusions regarding the effect of gender on faculty evaluations. Thus, this study adds to existing SEI literatures by not only adding in considerations of perceptions of faculty role congruity but also by taking a step back and testing for any potential gender and role congruity-based biases in the SEI forms themselves. By determining if there are any measurement biases between men and women instructors in the SEI forms themselves, these measurement biases can be accounted for to better compare the mean SEI scores of men and women instructors.

DATA

Student evaluation of instructors data as well as human resources data from a large, research-intensive, land-grant university in the Appalachian region of the United States were utilized in this study. See Chapter 1 for a complete description of the data cleaning, merging, and variable creation process. The data used in this chapter include quantitative SEI responses from

three semesters of evaluations: fall 2017, spring 2018, and fall 2018. The data includes student, course, and faculty information. In this study, analyses will utilize the quantitative SEI responses, instructor sex (coded as male, female with all others removed), and the gender dominance of the discipline (coded as women-dominated, men-dominated, and neutral).

The instructor sex and gender dominance of the discipline variables were used in conjunction to create role congruity variables, the full process for which can be found in Chapter 1. From the role congruity variables, three faculty congruity dummy variables were created such that each category of role congruity was set to be equal to one and all else was set to be missing. The dummy variables for the faculty groups allowed for easy selection of one group of faculty congruity at a time during the analyses described in the chapters that follow. The three faculty groups are Role Congruent which includes women in humanities and men in STEM, Role Incongruent which includes women in STEM, and men in humanities, and Role Neutral which includes women and men in disciplines that are not dominated by a particular gender.

The three faculty congruity dummy variables were used to create three separate datasets for analyses so that each congruity category could be analyzed separately. While comparing the differences in quantitative SEI scores across role congruity groups (e.g. how do role congruent scores compare to role incongruent scores) is useful, the main purpose of this study is to compare the SEI scores within a category of role congruity. For example, in this study, the scores of men instructors who teach in role congruent disciplines will be compared to the scores of women instructors who teach in role congruent disciplines. In this way, I will be able to compare the scores of women and men instructors who all experience the same level of discipline role congruity to one another to determine if there are or are not differences in SEI results.

Seven quantitative student evaluation questions were included in the analyses. These seven questions, presented in Table 1, were chosen because they were the only questions that were the asked on every SEI for the three semesters included in these analyses. Each of these questions was asked on a five-point Likert scale. Six of the seven questions are answered on a scale with the options of strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree. The question labeled Overall-Learning was answered using the following scale options: poor, fair, satisfactory, good, and excellent. A "response" is one student evaluation form from one student for one instructor for one course.

Table 1: Variables from Student Evaluations of Instruction

Variable Name	Student Evaluation Question	Response Count
Content-Related-Assignments	Course content was related to graded assignments	103,834
Content-Thought-Provoking	Course content was thought provoking	103,126
Material-Useful	The course materials were useful to course objectives	101,779
Positive-Learning-Environment	The instructor fostered a positive learning environment	101,594
Instructor-Organized	The instructor was well organized	101,344
Instructor-Feedback	The instructor provided helpful feedback	100,537
Overall-Learning	Overall my learning in this course was	103,390

METHODS AND ANALYSES

The finalized dataset was analyzed using STATA statistical software. Exploratory Factor Analyses (EFA) were the first statistical text completed. EFAs are used to determine which observed variables combine to measure the same latent variables/constructs. Observed variables are those that are actually measured or recorded, in this study the observed variables are the measures recorded on the student evaluation forms. Latent constructs are abstract concepts that the observed variables combine to measure but are not actually measured. Latent constructs are not directly measured or observed but rather they are inferred from the observed variables. For

example, quality of life would be considered a latent construct while observed variables such as wealth, occupation, housing, and more would be measured and combined to assess a person's quality of life. EFA models also include error terms on every observed variable which account for any bias to the measurement of the observed variables. EFAs are an exploratory model used to determine the structure of the relationship between the observed variables and latent constructs.

From the EFA model, Confirmatory Factor Analyses (CFA) for each faculty group were completed. CFAs are based on a hypothesized structure that can be determined by theory and/or EFAs. CFAs take EFAs a step further by testing the hypothesis that the structure of observed variables identified in the EFA does in fact measure the latent constructs. In this study, the CFA models test to see if there is a relationship between observed variables, SEI questions, and any identified latent constructs. The SEI forms themselves provide two likely latent constructs: the overall course quality and the quality of the instructor. These two latent concepts are likely to emerge from EFA and CFA testing because the questions on the SEI forms are arranged around these two themes thus indicating that the observed variables asked in each respective section are meant to measure their respective latent construct. The EFA and CFA testing will confirm or disconfirm that these two latent constructs are measured by the observed SEI measures in the ways in which the forms imply they do.

From the CFA, Multiple Indicators Multiple Causes (MIMIC) models were run. MIMIC models are a type of CFA in which the latent constructs as measured by the observed indicators are regressed on exogenous covariates, characteristics of the group in question which, in this study, is the gender of the instructors (Cao et al. 2019). MIMIC models are an ideal type of analysis for the examination of quantitative SEI data because they allow for the detection of

measurement equivalence, whether individual items are measured the same way across the groups that comprise the exogenous covariates, and whether these items exhibit differential item functioning (Diemer et al. 2019; Diemer and Li 2011). MIMIC models can detect if there are group differences in a measurement model and test whether each latent factor is measured in the same way across groups. Thus, MIMIC testing on SEI data will determine if the forms themselves are biased such that they lead to gender differences in student responses.

To take the analyses a step further, grouped Structural Equation Models (SEMs) were conducted to further test for measurement invariance and determine which, if any, observed variables are measured differently for the different gender groups. Measurement invariance analyses provide construct validity and psychometric support for the observed variables that measure latent constructs. The grouped SEMs determine if observed variables measure something different from one group to another and which specific observed variables are measured differently for the two gender groups. There are multiple steps in the grouped SEM process. In each step, different portions of the structural equation model are constrained or allowed to vary to test for a variety of differences between the groups. These steps must be completed before testing for differences in the means between the groups because they determine if there is measurement bias that needs to be accounted for prior to comparing means. It is crucial to detect and account for any measurement biases prior to comparing group means because any measurement biases present will affect the results. Thus, any differences detected in the means may not actually be due to mean differences but may actually be due to differences in measurement. For example, if a scale measures a ten-pound bag of apples as ten pounds but then measures a ten-pound bag of oranges as seven pounds, there is a problem with the measurement of the scale that is caused by the different groups being measured. If you do not correct for this

measurement error before comparing the groups, the group comparisons may be wildly inaccurate.

The first step is to test the model for configural invariance. Testing for configural invariance determines if the configuration of items loading onto latent constructs is the same across the groups. To test for configural invariance, a same form model is run in which there are no equality constraints placed on the coefficients of the observed variables, and the means are constrained and therefore not estimated. In this model, each group has the same form of observed variables loading on the latent constructs but the loadings do not need to be the same for each group (Acock 2013). The next step is to use an equal loadings model to test for metric invariance. Testing for metric invariance determines if the constructs manifest the same way in each group. In other words, it is testing if the slopes of the indicators on the factors are the same between the groups. To test for metric invariance, an equal loadings model is run in which the coefficients are now constrained to be equal across groups and the means again are constrained and therefore not estimated.

After the tests for configural and metric invariance, a post-hoc likelihood-ratio test is conducted to compare the performance of the two models. If the equal loadings model performs better than the same form model, testing for measurement invariance can proceed to a model that tests for equal loadings and equal error-variances and then a model that tests for equal intercepts. In the equal loadings and equal error-variances model, the loadings and measurement error variances are constrained to be equal across groups and the means again are constrained and therefore not estimated. In the equal intercepts model, the loadings and intercepts are constrained to be equal and means once again are constrained and therefore not estimated.

If it is determined that the equal loadings model (metric invariance) performs significantly worse than the same form model (configural invariance), a post estimation test can be conducted to determine which observed variables are measured differently for the two groups. The post estimation test "performs score tests (Lagrange multiplier tests) and Wald tests of whether parameters constrained to be equal across groups should be relaxed and whether parameters allowed to vary across groups could be constrained" (Anon 2021:149). If the post estimation test determines that *all* of the observed variables are measured differently, the next step is to conduct a same form equivalence model without the means constrained in order to compare the means between the two groups. In this same form equivalence model, the measurement intercepts are constrained to be equal across the groups, but the means are allowed to vary. The means are allowed to vary freely so that a means comparison can be done to determine which group scores are higher than the other.

If the equal loadings model performs worse than the same form model and the post estimation test determines that only *some* of the observed variables are measured differently, a partial invariant model can be conducted in order to compare the means between the groups. In a partial invariant model, the loadings of the observed variables which were determined to be measured the same across the groups are constrained to be equal while the observed variables which were determined to be measured differently and the means are allowed to vary. This model, therefore, allows what is measured differently between the groups to do so which allows for a more accurate comparison of the means between the groups.

The means comparison of the properly constrained models is the last step in the statistical analyses. Properly constraining the models to account for measurement invariance is a much more accurate way to compare the means between the groups because the constrained models are

accounting for measurement biases between the groups, in this case, men and women instructors. Thus, utilizing either the same form equivalence model or the partial invariant loadings model is a much more accurate way to compare the means between men and women instructors because the differences in measurement between the two gender groups are being accounted for whereas these measurement differences are not accounted for in, for example, a regression model. The results of this study will show if there are measurement biases and mean score differences in student evaluations based on students' perceptions of the gender and gender role congruity of their instructors.

RESULTS

The datasets included 99,545 role congruent responses, 51,684 role incongruent responses, and 25,076 role neutral responses. The analyses conducted utilized listwise deletion so any incomplete responses were automatically dropped from the analyses. For example, if an evaluation included responses for all but one of the questions included in the models, that evaluation was not included in the analyses. Table 2 describes the breakdown of the sample by faculty group and gender. A "response" is a single student evaluation from one student for one professor about one course.[5] The results for the exploratory factor analysis are presented first followed by a results section with the results of the CFA, MIMIC, and grouped SEM models for each of the instructor groups: role congruent, role incongruent, and role neutral.

[5] The responses are non-independent as the data include multiple responses for a single faculty member and, potentially, multiple responses by a single student. Clustered standard errors could be used to control for the non-independence but clustering the standard errors in structural equation modeling does not allow for tests of fit or model comparisons. Thus, though it is a limitation, the non-independence of responses is not controlled for in these analyses.

Table 2: SEI Response Counts by Faculty Group and Gender

Faculty Group	Men Responses	Women Responses	Total Responses
Role Congruent	33,312	21,624	54,936
Role Incongruent	14,929	12,975	27,904
Role Neutral	6,183	6,884	13,067

Exploratory Factor Analysis

Utilizing all of the available data, an exploratory factor analyses was completed to test if the observed variables load onto the latent constructs (unobserved variables) they were grouped into on the SEI forms: overall course quality (Overall) and instructor quality (Instructor). The check for bivariate normality for the observed variables on the Overall latent factor indicated that the data violate normality thus maximum likelihood mean-variance adjusted analyses could not be conducted. Therefore, exploratory factor analyses with maximum likelihood testing were completed. Results indicated strong positive correlations between the four observed variables for the latent concept Overall those being Content-Related-Assignments, Content-Thought-Provoking, Material-Useful, and Overall-Learning. The correlation matrix can be seen in Table 3. Factor testing revealed that there was a clear one-factor solution as all factor loadings were greater than 0.7 with a Cronbach's alpha of 0.8814 and item-rest correlations between 0.6638 and 0.8154. Factor loadings need to be greater than 0.3 and item-rest correlations need to be greater than 0.5, both of which are the case in this model indicating that this is a well-fit model.

Table 3: Correlation Matrix of Observed Variables for the Latent Concept Overall

	Content-Related-Assignments	Content-Thought-Provoking	Material-Useful	Overall-Learning
Content-Related-Assignments	1.0000			
Content-Thought-Provoking	0.6834	1.0000		
Material-Useful	0.7727	0.7258	1.0000	
Overall-Learning	0.5704	0.5994	0.6317	1.0000

The check for bivariate normality for the latent concept of Instructor would not run, therefore other normality checks were completed. Kaiser–Meyer–Olkin test results indicated a value of 0.746 with a determinate of the correlation matrix of 0.179 thus indicating that the data are acceptable for factor analyses. Exploratory factor analyses with maximum likelihood testing showed strong positive correlations between the three observed variables for the latent concept Instructor those being Positive-Learning-Environment, Instructor-Organized, and Instructor-Feedback. The correlation matrix can be seen in Table 4. Factor testing revealed that there was a clear one-factor solution as all factor loadings were greater than 0.81 with a Cronbach's alpha of 0.8879 and item-rest correlations between 0.7647 and 0.8018. The rotated factor analysis, oblique rotation, and orthogonal rotation all revealed the same factor loading matrix thus indicating that the unrotated one-factor solution is appropriate.

Table 4: Correlation Matrix of Observed Variables for the Latent Concept Instructor

	Positive-Learning-Environment	Instructor-Organized	Instructor-Feedback
Positive-Learning-Environment	1.0000		
Instructor-Organized	0.7030	1.0000	
Instructor-Feedback	0.7526	0.7294	1.0000

Thus, the EFA determined that there was one model with two latent concepts, Instructor and Overall, which were measured by the observed variables. The latent concepts were named based on the Student Evaluation of Instructors forms which subdivided the questions into these two categories. The EFA confirms the general university organization of the SEIs (i.e., a set of questions regarding the evaluations of the instructor and a set of questions measure evaluations of the course). Now that the two factors are identified and deemed statistically appropriate,

testing by instructor group can proceed with CFA, MIMIC, and Grouped SEM models. The results of these tests follow and are organized by instructor role congruity group.

Role Congruent Faculty

Confirmatory Factor Analyses

A two-factor Confirmatory Factor Analysis was conducted with both latent constructs – Overall and Instructor - and their identified measured variables combined into one model. The model fit was sufficient with a CFI (0.968) slightly above and TLI (0.948) slightly below the 0.95 cutoff point ($\chi 2(13)=8651.78$, $p<0.001$). The SRMR (0.043) was below the 0.06 cutoff, but the RMSEA (0.110) was above the 0.05 cutoff point. The rho reliability (0.94) further indicates that the model fit is acceptable, though as indicated by the fit statistics it could be improved. The model modification indices further indicate that there are ways in which the model could be improved with the largest expected parameter change coming from adding a path from the observed variable of Overall-Learning to the latent construct Instructor (EPC=0.9146).

A one-factor CFA of just the latent concept Instructor with the added path from Overall-Learning was conducted to test the appropriateness of adding this path to the two-factor model. The results of the one-factor CFA for Instructor with the added path from Overall-Learning indicate that this additional path is very appropriate ($\chi 2$ (2)=49.18, $p<0.001$; CFI=1.000, TLI=0.999, SRMR=0.003, RMSEA=0.020). Due to the exceptional model fit, modification indices were not explored further. The path from Overall-Learning to Instructor was then added to the two-factor CFA model. The two-factor model now included a path from Overall-Learning to both latent constructs (Overall and Instructor). The latent construct Instructor was still also measured by the observed variables Positive-Learning-Environment, Instructor-Organized, and Instructor-Feedback, and the latent construct Overall was still also measured by the observed variables Content-Related-Assignments, Content-Thought-Provoking, and Material-Useful. The

addition of the path from Overall-Learning to Instructor drastically improved the two-factor CFA model fit ($\chi 2$ (12)=1060.95, p<0.001; CFI=0.996, TLI=0.993, SRMR=0.011, RMSEA=0.040) and while there were modification indices, due to the exceptional fit and lack of theoretical reason behind adding any additional paths, further possible paths were not explored.

The final two-factor CFA model is illustrated in Figure 1. The variables that combined to measure the latent concept Overall were: Content-Related-Assignments, Content-Thought-Provoking, Material-Useful, and Overall-Learning. These four variables measure the "overall" quality of the course by asking about things such as the content, materials, and overall student learning in the course. The variables that combined to measure the latent concept Instructor were: Positive-Learning-Environment, Instructor-Organized, Instructor-Feedback, and Overall-Learning. These variables measure the quality of the "instructor" of the course by asking about things such as the kind of learning environment they create, their organization, the feedback they give, and the overall learning of the student in the course. Overall-Learning contributed to the measurement of both latent constructs indicating that students' perceptions of their overall learning in the course affects their evaluations of both the overall quality of the course and the quality of the instructor specifically. The "e"s in the figure represent the error term for each observed variable.

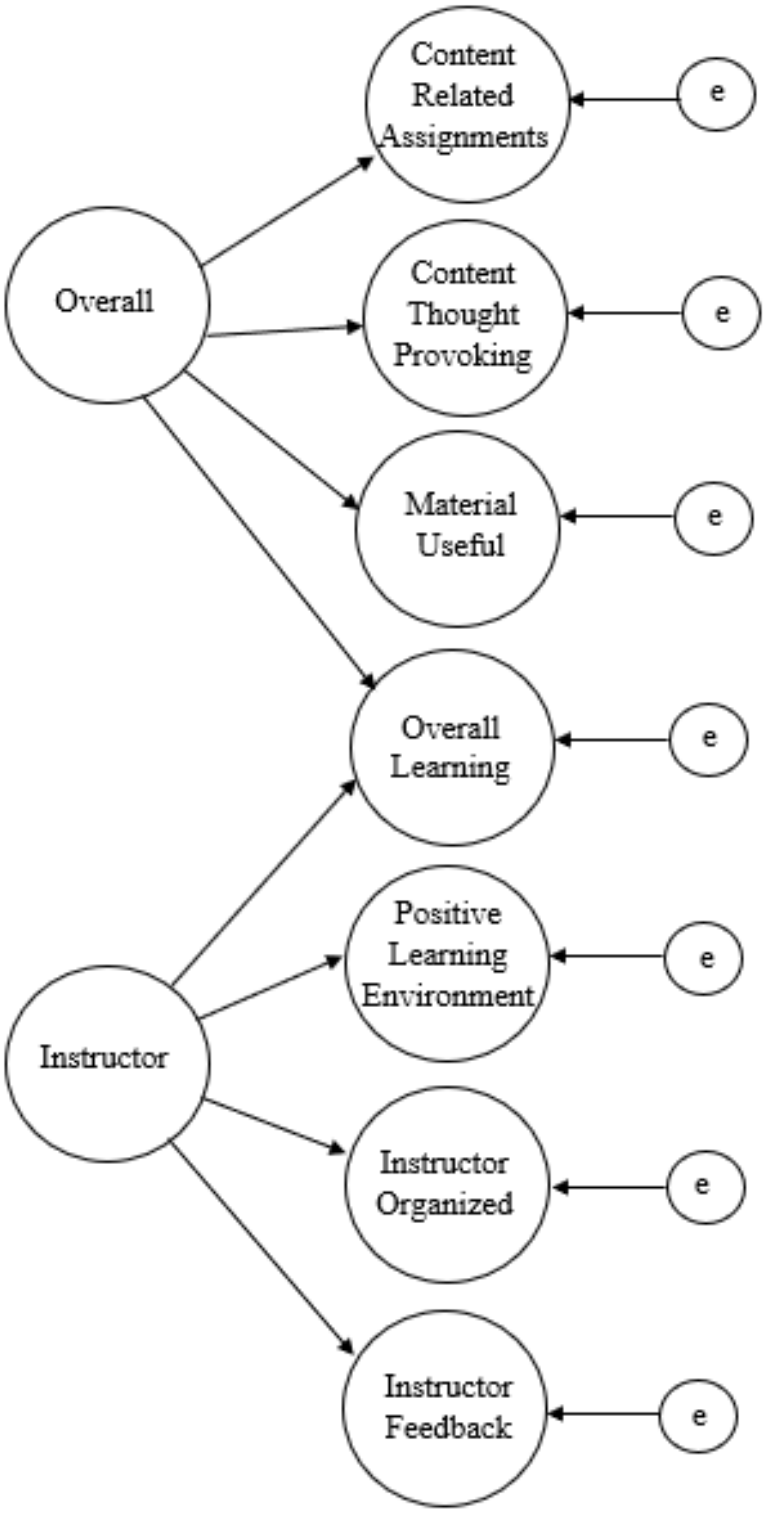


Figure 1: Diagram of the two-factor CFA Model. Overall and Instructor are latent concepts measured by the observed variables of Content-Related-Assignments, Content-Thought-Provoking, Material-Useful, Overall-Learning, Positive-Learning-Environment, Instructor-Organized, and Instructor-Feedback.

Multiple-Indicators Multiple-Causes Model

The well-fitting two-factor CFA was then used in a MIMIC model with instructor gender added as an exogenous covariate on the latent constructs which can be seen in Figure 2. The MIMIC model did not fully meet fit parameters ($\chi 2(18)=37942.77$, $p<0.001$; CFI=0.860, TLI=0.783, SRMR=0.304, RMSEA=0.196). The MIMIC model indicates that there are differences in how the observed variables are measured depending upon the gender of the instructor for both the Overall (0.0636, $p<0.001$) and Instructor (0.1260, $p<0.001$) latent constructs. Since group measurement differences between the groups were identified, a grouped structural equation model was conducted to determine more specific differences in measurement between the groups.

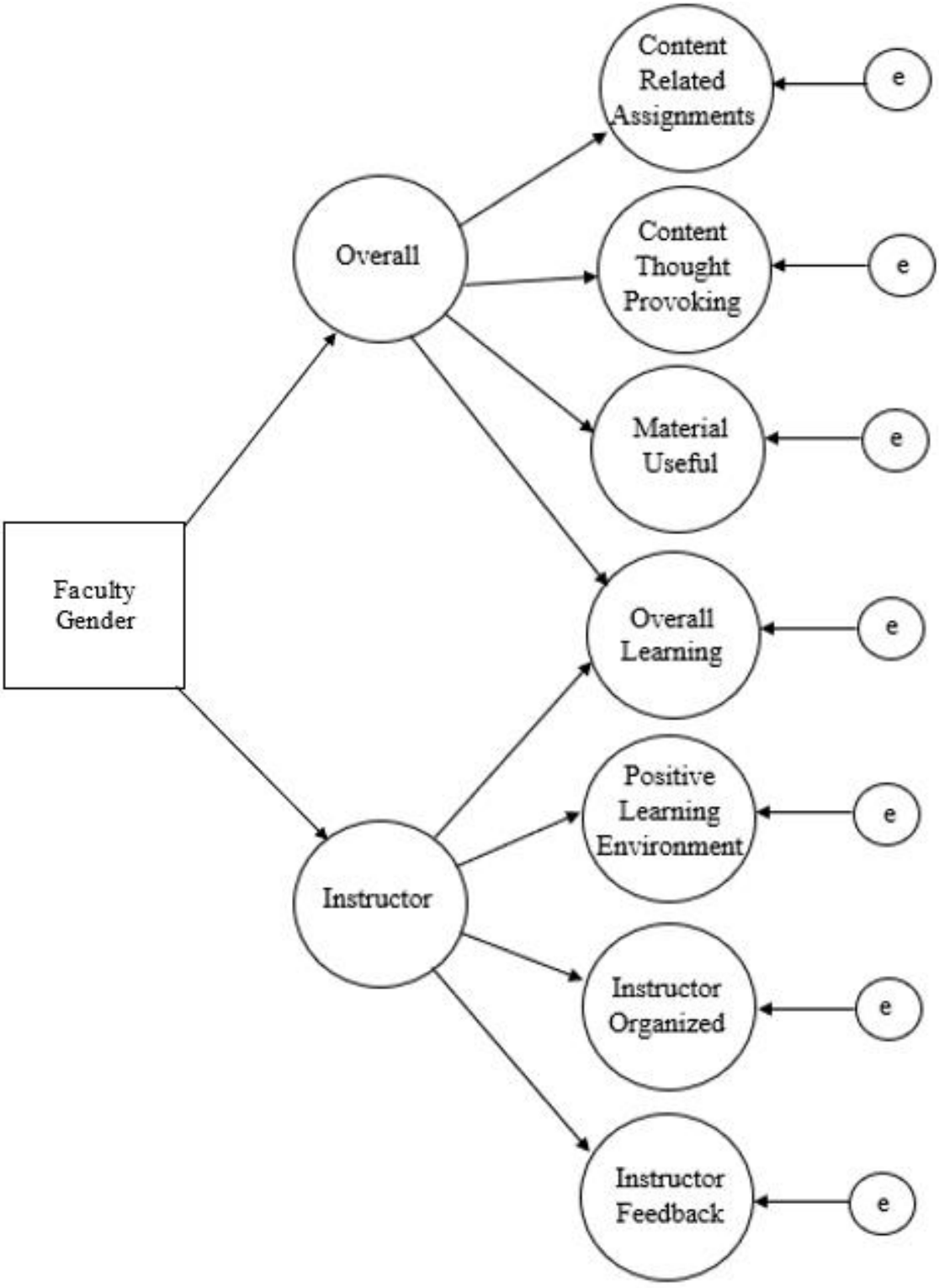


Figure 2: Diagram of the MIMIC Model. Faculty gender is an exogenous covariate on Overall and Instructor which are latent concepts measured by the observed variables of Content-Related-Assignments, Content-Thought-Provoking, Material-Useful, Overall-Learning, Positive-Learning-Environment, Instructor-Organized, and Instructor-Feedback.

Grouped Structural Equation Models

A grouped SEM was conducted utilizing the CFA model from Figure 1 and grouped based on instructor gender: role congruent men vs. role congruent women. To test for configural invariance, a same form equivalence model was conducted in which the means of the latent concepts were set to equal zero but there were no constraints placed on the groups. The results of this SEM model and the subsequent models are in Table 5. The loadings were all substantial and statistically significant and the model fit statistics all indicated that the model fit well ($\chi 2(24)=1134.02$, $p<0.001$; CFI=0.996, TLI=0.993, SRMR=0.011, RMSEA=0.041).

To test for metric invariance, an equal loadings model was conducted in which the means of the latent concepts were still set to equal zero and the measurement coefficients are constrained to be equal across the groups. The loadings were once again all substantial and statistically significant and the model fit statistics all indicated that the model fit well ($\chi 2(30)=1273.25$, $p<0.001$; CFI=0.995, TLI=0.994, SRMR=0.016, RMSEA=0.039).

A likelihood-ratio test was then used to compare the two models. The likelihood ratio test indicated that the equal loadings model performs statistically significantly worse than the same form model ($\chi 2(6)=139.22$, $p<0.001$). Thus, we should not constrain the loadings to be equal. This means that there are statistically significant differences between women and men in the meaning of the latent variables Instructor and Overall when measured with the observed variables used in these models. Because the overall model fit of the equal loadings model is worse than the same form equivalence model this means that at least one loading is not equivalent across the groups. Metric invariance is therefore *not* supported.

Table 5: Comparison of the Three Grouped Structural Equation Models for Role Congruent Faculty

	Same Form Equivalence				Equal Loadings Model			
	Men (*N*=33,312)		Women (*N*=21,624)		Men (*N*=33,312)		Women (*N*=21,624)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	a	0.84***	a	**0.86***	a	0.84***	a	0.86***
Content-Thought-Provoking	1.05***	0.79***	1.09***	**0.83***	1.06***	0.80***	1.06***	0.83***
Material-Useful	1.15***	0.90***	1.18***	**0.90***	1.17***	0.90***	1.17***	0.90***
Overall-Learning	0.47***	**0.31***	0.38***	0.26***	0.43***	0.28***	0.43***	0.29***
Instructor								
Positive-Learning-Environment	1.37***	**0.85***	1.23***	0.85***	1.30***	0.85***	1.30***	0.86***
Instructor-Organized	1.39***	0.82***	1.34***	**0.83***	1.36***	0.82***	1.36***	0.82***
Instructor-Feedback	1.62***	0.88***	1.53***	**0.88***	1.57***	0.88***	1.57***	0.88***
Overall-Learning	a	0.52***	a	**0.53***	a	0.54***	a	0.50***
Mean Overall	a	a	a	a	a	a	a	a
Mean Instructor	a	a	a	a	a	a	a	a
R^2	0.979		0.982		0.979		0.982	
$\chi 2$	*df*=24,1134.02***				*df*=30, 1273.25***			
CFI	0.996				0.995			
RMSEA	0.041				0.039			

B=unstandardized, *β*=standardized
a Not reported because of constraints
*p<0.05; **p<0.01; ***p<0.001
Bold text indicates the higher loading between the two gender groups
Note: p-values indicated that the loadings are significant **not** that there are differences between the groups

A post estimation test indicates that all of the observed variables in the model differ significantly between men and women in the level of importance they carry in their measurement of the latent concepts. The output of the post-estimation test shows significant chi-squared values for all of the variables in the model which can be seen in Table 6. This means that all of the variables differ significantly on their levels of importance for men and women.

Table 6: Test of Group Invariance of Parameters

Latent Variables	Observed Variables	Score Test		
		χ^2	*df*	$p>\chi^2$
Overall	Content-Related-Assignments	15.748	1	0.0001
	Content-Thought-Provoking	10.900	1	0.0010
	Material-Useful	6.122	1	0.0134
	Overall-Learning	29.642	1	0.0000
Instructor	Positive-Learning-Environment	58.108	1	0.0000
	Instructor-Organized	47.201	1	0.0000
	Instructor-Feedback	7.332	1	0.0068
	Overall-Learning	8.139	1	0.0043

According to the standardized loadings[6] on the same form model (Table 5), Overall-Learning on the latent concept Overall and Positive-Learning-Environment on the latent concept Instructor weigh more in the measurement of men's scores than women's scores. A variable "weighing more" means that there is a greater strength of association between the observed variable and the latent construct. All other variables - Content-Related-Assignments, Content-Thought-Provoking, Material-Useful on the latent concept Overall and Instructor-Organized, Instructor-Feedback, and Overall-Learning on the latent concept Instructor - weigh more in the measurement of women's scores than men's scores. When standardized coefficients were the same rounded to two decimals, the full reported value was considered when selecting which was higher. Thus, Overall-Learning on Overall and Overall and Positive-Learning-Environment on Instructor have a greater strength of association with the measurement of their respective latent constructs for men than for women. All other variables have a greater strength of association with the measurement of their respective latent constructs for women than for men.

[6] For group comparisons, typically the unstandardized loadings are compared as they indicate the form of the relationship in this case, what is the actual difference in scores between men and women, while the standardized coefficients indicate the strength of the relationship between the observed variable and the latent construct. Due to the same form model being where the analyses need to stop, the standardized coefficients are compared so that all variables can be compared as opposed to missing some comparisons due to model constraints. This means that the group comparisons are saying that the variable in question has a stronger relationship with the latent construct for one group as compared to the strength of the relationship between the observed and latent variable for the other group.

All the observed variables were shown to be different across groups, thus we can only assume same form equivalence. This means that we can only compare the score means between men and women if same form equivalence is specified and the means are allowed to vary. Thus, a same form equivalence model was conducted with unconstrained means. This model and the output is identical to the same form equivalence model presented in Table 5 except that the means were allowed to vary thus allowing for the means to be compared between the groups. The results indicate that for the latent concept of Overall women are rated slightly higher than men (0.0973, $p<0.001$), and women are also rated slightly higher than men on the latent concept of Instructor (0.1611, $p<0.001$). These results indicate that there are score differences on SEI forms based on instructor gender and role congruity such that role congruent women instructors are rated slightly higher than their role congruent man peers on both latent concepts: Overall and Instructor. While there are still problems with this means comparison because our latent variables have different meanings for men and women as indicated by the lack of metric invariance, these results are better than a traditional *t*-test or even a coefficient in a regression because they do account for measurement invariance between the groups.

Role Incongruent Faculty

Confirmatory Factor Analyses

Once again, a two-factor Confirmatory Factor Analysis (Figure 1) was conducted with both latent constructs and their identified measured variables combined into one model. The model fits well with a CFI (0.972) and TLI (0.955) both above the 0.95 cutoff point ($\chi 2(13)=386806$, $p<0.001$). The SRMR (0.043) was below the 0.06 cutoff, but the RMSEA (0.103) was above the 0.05 cutoff point. The rho reliability (0.94) further indicates that the model fit is acceptable, though as indicated by the fit statistics it could be improved. The model modification indices further indicate that there are ways in which the model could be improved

with the largest expected parameter change coming from adding a path from the observed variable of Overall-Learning to the latent construct Instructor (EPC=0.8620).

Much like with Role Congruent faculty, a one-factor CFA of just Instructor was conducted to test the appropriateness of adding a path from Overall-Learning to Instructor. The results of the one-factor CFA indicate that this additional path is very appropriate ($\chi 2(2)=17.67$, p =0.0001; CFI=1.000, TLI=0.999, SRMR=0.003, RMSEA=0.016). Due to the exceptional model fit, modification indices were not explored further. The path from Overall-Learning to Instructor was then added to the two-factor CFA model. The two-factor CFA model fit improved drastically ($\chi 2(12)=504.81$, p <0.001; CFI=0.996, TLI=0.994, SRMR=0.010, RMSEA=0.038) and while there were modification indices, due to the exceptional fit and lack of theoretical reason behind adding any additional paths, further possible paths were not explored. The final two-factor model can be seen in Figure 1.

Multiple-Indicators Multiple-Causes Model

The well-fitting two-factor CFA was then used in a MIMIC model (Figure 2) with instructor gender added as an exogenous covariate on the latent constructs. The MIMIC model did not fully meet fit parameters ($\chi 2(18)=19578.42$, p<0.001; CFI=0.858, TLI=0.780, SRMR=0.305, RMSEA=0.197). The MIMIC model indicates that there are differences in how the observed variables are measured depending upon the gender of the instructor for both the Overall (-0.0795, p<0.001) and Instructor (-0.0808, p<0.001) latent constructs. Since group measurement differences between the groups were identified, a grouped structural equation model was conducted to determine more specific differences in measurement between the groups.

Grouped Structural Equation Models

A grouped SEM was conducted utilizing the CFA model from Figure 1 and grouped based on instructor gender. The output for all of the models conducted is presented in Table 8. To test for configural invariance, a same form equivalence model was conducted in which the means of the latent concepts were set to equal zero but there were no constraints placed on the groups. The loadings were all substantial and statistically significant and the model fit statistics all indicated that the model fit well ($\chi 2(24)=553.48$, $p<0.001$; CFI=0.996, TLI=0.993, SRMR=0.011, RMSEA=0.040). To test for metric invariance, an equal loadings model was conducted in which the means of the latent concepts were still set to equal zero and the measurement coefficients are constrained to be equal across the groups. The loadings were once again all substantial and statistically significant and the model fit statistics all indicated that the model fit well ($\chi 2(30)=593.53$, $p<0.001$; CFI=0.996, TLI=0.994, SRMR=0.013, RMSEA=0.037).

A likelihood-ratio test was then used to compare the two models. The likelihood ratio test indicated that the equal loading model performs statistically significantly worse than the same form model ($\chi 2(6)=40.05$, $p<0.001$). Thus, we should not constrain the loadings to be equal. This means that there are statistically significant differences between women and men in the meaning of the latent variables Instructor and Overall when measured with the observed variables used in these models. Because the overall model fit of the equal loadings model is worse than the same form equivalence model this means that at least one loading is not equivalent across the groups. Metric invariance is therefore *not* supported.

Table 7: Comparison of the Three Grouped Structural Equation Models for Role Incongruent Faculty

	Same Form Equivalence				Equal Loadings				Partial Invariant Loadings			
	Men (*N*=14,929)		Women (*N*=12,975)		Men (*N*=14,929)		Women (*N*=12,975)		Men (*N*=14,929)		Women (*N*=12,975)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall												
Content-Related-Assignments	a	0.85***	a	0.85***	a	0.85***	a	0.85***	1.16	0.85***	1.16	0.85***
Content-Thought-Provoking	1.06***	0.83***	1.06***	0.80***	1.06***	0.83***	1.06***	0.80***	1.22	0.83***	1.22	0.80***
Material-Useful	1.15***	0.90***	1.16***	0.80***	1.16***	0.90***	1.16***	0.90***	1.34	0.90***	1.34	0.90***
Overall-Learning	0.37***	0.25***	0.51***	0.34***	0.43***	0.30***	0.43***	0.29***	0.47	0.28***	0.54	0.31***
Instructor												
Positive-Learning-Environment	1.28***	0.85***	1.46***	0.87***	1.35***	0.85***	1.35***	0.87***	0.94	0.85***	0.94	0.87***
Instructor-Organized	1.32***	0.82***	1.51***	0.84***	1.40***	0.82***	1.40***	0.84***	0.97	0.82***	0.97	0.84***
Instructor-Feedback	1.52***	0.86***	1.74***	0.89***	1.61***	0.86***	1.61***	0.89***	1.12	0.87***	1.12	0.89***
Overall-Learning	a	0.53***	a	0.47***	a	0.49***	a	0.51***	0.69	0.50***	0.69	0.50***
Mean Overall	a	a	a	a	a	a	a	a	a	a	a	a
Mean Instructor	a	a	a	a	a	a	a	a	a	a	a	a
R^2	0.980		0.981		0.980		0.981		0.980		0.981	
χ2	*df*=24, 553.48***				*df*=30, 593.53***				*df*=27, 571.90***			
CFI	0.996				0.996				0.996			
RMSEA	0.040				0.037				0.038			

B=unstandardized, *β*=standardized
a Not reported because of constraints
*p<0.05; **p<0.01; ***p<0.001

A post estimation test indicates that one of the observed variables in the model differs significantly between men and women in the level of importance it carries in the measurement of the latent concepts. The results are presented in Table 9. There are significant chi-squared values for Overall-Learning (21.602, $p<0.001$) on the latent concept Overall. This means that only the variable Overall-Learning when it is measuring the latent concept of Overall differs significantly on its level of importance for men and women. According to the partial invariant loadings model, Overall-Learning on the latent concept Overall carried more weight for women instructors (0.54) than for men instructors (0.47). This means that Overall-Learning has a greater effect on the Overall score of women than men.

Table 8: Test of Group Invariance of Parameters

Latent Variables	Observed Variables	Score Test		
		$\chi 2$	*df*	$p>\chi 2$
Overall	Content-Related-Assignments	0.458	1	0.4987
	Content-Thought-Provoking	1.793	1	0.1805
	Material-Useful	0.139	1	0.7093
	Overall-Learning	21.602	1	0.0000
Instructor	Positive-Learning-Environment	1.037	1	0.3085
	Instructor-Organized	0.085	1	0.7705
	Instructor-Feedback	0.037	1	0.8478
	Overall-Learning	3.004	1	0.0830

Since only one variable was shown to be different across groups, a partial invariant loadings model can be run in order to compare the mean scores between men and women. In the partial invariant loadings model, the loadings for all of the variables which were determined to not be measured differently were constrained to be equal while the loading for the one variable that was determined to be different was allowed to vary. Additionally, the means were allowed to vary. The loadings for the partially invariant model were all positive but not statistically significant. The model fit statistics all indicated that the model fit well ($\chi 2(27)=571.90$, $p<0.001$; CFI=0.996, TLI=0.994, SRMR=0.012, RMSEA=0.038).

Given the good fit statistics, the partially invariant model can be used to test for differences in the means between the groups, those being men and women. The results of the partially invariant means comparison model can be seen in table 10. The model fit well ($\chi 2(32)=963.14$, $p<0.001$; CFI=0.993, TLI=0.991, SRMR=0.012, RMSEA=0.046) and the unstandardized loadings[7] were all positive, though not statistically significant. The results indicate that there are not statistically significant differences between the scores of role incongruent men and women on the latent concepts of Overall (-0.1211, $p=0.956$) or Instructor (-0.1193, $p=0.918$). These results indicate that when the model is properly constrained for measurement differences, there are *not* score differences on SEI forms between role incongruent men and women.

[7] The loadings and significance are different between the unstandardized and standardized solutions because the model is invariant and the standardized solution confounds the invariance in the form of the relationship with group differences in the standardized deviations.

Table 9: Partial Invariant Loadings Model Means Comparison

	Partial Invariant Loadings Means Comparison			
	Men (*N*=14,929)		Women (*N*=12,975)	
	B	*β*	*B*	*β*
Overall				
Content-Related-Assignments	1.02	0.85***	1.02	0.85***
Content-Thought-Provoking	1.09	0.83***	1.09	0.80***
Material-Useful	1.18	0.90***	1.18	0.90***
Overall-Learning	0.41	0.28***	0.49	0.31***
Instructor				
Positive-Learning-Environment	1.04	0.85***	1.04	0.87***
Instructor-Organized	1.08	0.82***	1.08	0.84***
Instructor-Feedback	1.24	0.86***	1.24	0.89***
Overall-Learning	0.77	0.50***	0.77	0.50***
Mean Overall	a	a	-0.12	a
Mean Instructor	a	a	-0.12	a
R^2	0.980		0.981	
χ2	*df*=32, 963.14***			
CFI	0.993			
RMSEA	0.043			

B=unstandardized, *β*=standardized
a Not reported because of constraints
*p<0.05; **p<0.01; ***p<0.001

Role Neutral Faculty

Confirmatory Factor Analyses

For the role neutral faculty, analyses once again began with a two-factor Confirmatory Factor Analysis (Figure 1) with both latent constructs and their identified measured variables combined into one model. The model fits well with a CFI (0.970) and TLI (0.952) both above the 0.95 cutoff point ($\chi 2(13)=2001.42$, $p<0.001$). The SRMR (0.035) was below the 0.06 cutoff, but the RMSEA (0.108) was above the 0.05 cutoff point. The rho reliability (0.94) further indicates that the model fit is acceptable, though as indicated by the fit statistics it could be improved. The model modification indices further indicate that there are ways in which the model could be

improved with the largest expected parameter change coming from adding a path from the observed variable of Overall-Learning to the latent construct Instructor (EPC=1.003).

Much like with Role Congruent and Role Incongruent faculty, a one-factor CFA of just Instructor was conducted to test the appropriateness of adding a path from Overall-Learning to Instructor. The results of the one-factor CFA indicate that this additional path is very appropriate ($\chi 2(2)=2.45$, $p=0.2940$; CFI=1.000, TLI=1.000, SRMR=0.001, and RMSEA=0.004). Due to the exceptional model fit, modification indices were not explored further. The path from Overall-Learning to Instructor was then added to the two-factor CFA model. The two-factor CFA model fit improved drastically ($\chi 2(12)=272.68$, $p<0.001$; CFI=0.996, TLI=0.993, SRMR=0.011, RMSEA=0.041) and while there were modification indices, due to the exceptional fit and lack of theoretical reason behind adding any additional paths, further possible paths were not explored. The final two-factor model can be seen in Figure 1.

Multiple-Indicators Multiple-Causes Model

The well-fitting two-factor CFA was then used in a MIMIC model (Figure 2) with instructor gender added as an exogenous covariate on the latent constructs. The MIMIC model did not fully meet fit parameters ($\chi 2(18)=41425.22$, $p<0.001$; CFI=0.849, TLI=0.766, SRMR=0.320, RMSEA=0.207). The MIMIC model indicates that there are differences in how the observed variables are measured depending upon the gender of the instructor for both the Overall (-0.042, $p<0.001$) and Instructor (-0.066, $p<0.001$) latent constructs. Since group measurement differences between the groups were identified, a grouped structural equation model was conducted to determine more specific differences in measurement between the groups.

Grouped Structural Equation Models

A grouped SEM was conducted utilizing the CFA model from Figure 1 and grouped based on instructor gender. The results of all of the SEM models run are in Table 11. To test for configural invariance, a same form equivalence model was conducted in which the means of the latent concepts were set to equal zero but there were no constraints placed on the groups. The loadings were all substantial and statistically significant and the model fit statistics all indicated that the model fit well ($\chi 2(24)=315.32$, $p<0.001$; CFI=0.996, TLI=0.992, SRMR=0.011, RMSEA=0.043). To test for metric invariance, the means of the latent concepts were still set to equal zero and the measurement coefficients are constrained to be equal across the groups. The loadings were once again all substantial and statistically significant and the model fit statistics all indicated that the model fit well ($\chi 2(30)=336.89$, $p<0.001$; CFI=0.995, TLI=0.994, SRMR=0.015, RMSEA=0.040).

A likelihood-ratio test was then used to compare the two models. The likelihood ratio test indicated that the equal loadings model performs statistically significantly worse than the same form model ($\chi 2(6)=21.57$, $p=0.0014$). Thus, we should not constrain the loadings to be equal. This means that there are statistically significant differences between women and men in the meaning of the latent variables Instructor and Overall when measured with the observed variables used in these models. Because the overall model fit of the equal loadings model is worse than the same form equivalence model this means that at least one loading is not equivalent across the groups. Metric invariance is therefore *not* supported.

Table 10: Comparison of the Three Grouped Structural Equation Models for Role Neutral Faculty

	Same Form Equivalence				Equal Loadings				Partial Invariant Loadings			
	Men (*N*=6,183)		Women (*N*=6,884)		Men (*N*=6,183)		Women (*N*=6,884)		Men (*N*=6,183)		Women (*N*=6,884)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall												
Content-Related-Assignments	a	0.84***	a	0.85***	a	0.84***	a	0.85***	0.99	0.84***	0.95	0.85***
Content-Thought-Provoking	1.02***	0.80***	1.08***	0.81***	1.05***	0.81***	1.05***	0.81***	1.00	0.80***	1.03	0.81***
Material-Useful	1.13***	0.90***	1.18***	0.90***	1.16***	0.90***	1.16***	0.90***	1.12	0.90***	1.12	0.90***
Overall-Learning	0.46***	0.30***	0.51***	0.32***	0.49***	0.32***	0.49***	0.32***	0.47	0.32***	0.47	0.32***
Instructor												
Positive-Learning-Environment	1.22***	0.84***	1.32***	0.85***	1.28***	0.84***	1.28***	0.84***	0.93	0.84***	0.93	0.85***
Instructor-Organized	1.31***	0.81***	1.34***	0.82***	1.33***	0.80***	1.33***	0.82***	0.99	0.81***	0.95	0.82***
Instructor-Feedback	1.50***	0.88***	1.61***	0.90***	1.57***	0.88***	1.57***	0.90***	1.13	0.88***	1.13	0.90***
Overall-Learning	a	0.53***	a	0.54***	a	0.51***	a	0.55***	0.72	0.51***	0.72	0.55***
Mean Overall	a	a	a	a	a	a	a	a	a	a	a	a
Mean Instructor	a	a	a	a	a	a	a	a	a	a	a	a
R^2	0.975		0.981		0.975		0.981		0.975		0.981	
χ2	*df*=24, 315.32***				*df*=30, 336.89***				*df*=25, 318.79***			
CFI	0.996				0.995				0.996			
RMSEA	0.043				0.040				0.042			

B=unstandardized, *β*=standardized
a Not reported because of constraints
*p<0.05; **p<0.01; ***p<0.001

A post estimation test indicates that three of the observed variables in the model differ significantly between men and women in the level of importance they carry in their measurement of the latent concepts. The output which can be seen in Table 12 shows significant chi-squared values for Content-Related-Assignments (9.542, p=0.0020) and Content-Thought-Provoking (6.439, p=0.0112) on the latent concept Overall and Instructor-Organized (6.124, p=0.0133) on the latent concept Instructor. This means that these three variables (Content-Related-Assignments, ConThghtPro18, and Instructor-Organized differ significantly on their levels of importance for men and women.

Table 11: Test of Group Invariance of Parameters

Latent Variables	Observed Variables	Score Test		
		$\chi 2$	*df*	$p>\chi 2$
Overall	Content-Related-Assignments	9.542	1	0.0020
	Content-Thought-Provoking	6.439	1	0.0112
	Material-Useful	0.730	1	0.3928
	Overall-Learning	0.273	1	0.6016
Instructor	Positive-Learning-Environment	2.978	1	0.3081
	Instructor-Organized	6.124	1	0.0844
	Instructor-Feedback	1.261	1	0.0133
	Overall-Learning	1.039	1	0.2614

According to the partial invariant loadings model, Content-Related-Assignments on the latent concept Overall and Instructor-Organized on the latent concept Instructor matter more for the measurement of men (0.99, 0.99) than for women (0.95, 0.95). Content-Thought-Provoking on the latent concept Overall matters more for the measurement of women's scores (1.03) than for men's scores (1.00). This means that Content-Related-Assignments and Instructor-Organized have greater effects on the Overall and Instructor scores for men than for women while Content-Thought-Provoking has a greater effect on the Overall score for women than for men.

Since only three variables were shown to be different across groups, a partial invariant loadings model can be run in order to compare the mean scores between men and women. In the partial invariant loadings model, the loadings for the four variables which were determined to not be measured differently were constrained to be equal while the loadings for the three variables that were determined to be different were allowed to vary. The loadings for the partially invariant model were all positive but not statistically significant. The model fit statistics all indicated that the model fit well ($\chi 2(25)=318.79$, $p<0.001$; CFI=0.996, TLI=0.993, SRMR=0.012, RMSEA=0.042).

Given the good fit statistics, the partially invariant model can be used to test for differences in the means between the groups, those being men and women. The partially invariant means comparison model results are presented in Table 13. The model fit well ($\chi 2(30)=351.30$, $p<0.001$; CFI=0.995, TLI=0.993, SRMR=0.012, RMSEA=0.040) and the unstandardized loadings were all positive, though not statistically significant. Furthermore, the results indicate that there is not a statistically significant difference between the scores of role neutral men and women on the latent concepts of Overall (-0.0606, $p=0.972$) or Instructor (-0.1242, $p=0.979$). These results indicate that when the model is properly constrained for measurement differences, there are *not* score differences on SEI forms based on instructor gender and role neutrality.

Table 12: Partial Invariant Loadings Model Means Comparison

	Partial Invariant Loadings Means Comparison			
	Men (*N*=14,929)		Women (*N*=12,975)	
	B	*β*	*B*	*β*
Overall				
Content-Related-Assignments	1.14	0.84***	1.10	0.85***
Content-Thought-Provoking	1.16	0.80***	1.19	0.82***
Material-Useful	1.29	0.90***	1.29	0.90***
Overall-Learning	0.55	0.32***	0.55	0.32***
Instructor				
Positive-Learning-Environment	0.89	0.84***	0.89	0.85***
Instructor-Organized	0.95	0.81***	0.91	0.81***
Instructor-Feedback	1.09	0.88***	1.09	0.90***
Overall-Learning	0.69	0.51***	0.69	0.55***
Mean Overall	a	a	-0.06	a
Mean Instructor	a	a	-0.12	a
R^2	0.975		0.981	
$\chi 2$	*df*=30, 351.30			
CFI	0.995			
RMSEA	0.040			

B=unstandardized, *β*=standardized
a Not reported because of constraints
*p<0.05; **p<0.01; ***p<0.001

DISCUSSION

This study illustrates the importance of determining the appropriateness of a measurement model prior to testing for SEI differences between men and women. In each of the three faculty groups, variables were determined to be measured differently depending on the gender of the instructor. This means that a simple comparison of means or regression test on the unconstrained model would have been biased by the measurement errors of the model and therefore provide unreliable results. In this study, the proper constraints were added to the three models so that appropriate comparisons of the mean scores of men and women instructors could be compared.

For perceived role congruent faculty, all variables in the model were determined to be measured differently between the two gender groups. Overall-Learning on the latent construct Overall and Positive-Learning-Environment on the latent construct Instructor carry more weight in the score measurement for men than for women while all other variables (Content-Related-Assignments, Content-Thought-Provoking, Material-Useful on the latent concept Overall and Instructor-Organized, Instructor-Feedback, and Overall-Learning on the latent concept Instructor) carry more weight in the measurement for women than for men.

For perceived role incongruent instructors, only one variable was shown to weigh differently in the scoring of men and women. Overall-Learning carried more weight in the measurement of the latent concept Overall for women than for men instructors. Overall-Learning had a greater weight in the measurement of men's scores than women's scores for perceived role congruent faculty but for role incongruent faculty, Overall-Learning had a greater effect on the measurement of Overall for women than for men. This indicates that there is something in particular about this specific question which is affected by the gender and perceived role (in)congruity of the instructor being evaluated. The SEI scores of men instructors in role congruent and the SEI scores of women instructors in perceived role incongruent disciplines are more affected by students' perceptions of their overall learning than the scores of women in perceived role congruent and men in perceived role incongruent disciplines. Taken together, these results indicate that students' perceptions of their overall learning in a course has a greater effect on the evaluations of instructors in man-dominated disciplines such as STEM fields than on women-dominated or role-neutral disciplines. In science, math, and other traditionally masculine fields, instructors need to ensure that their students feel like they have learned in the course in order to receive higher student evaluation scores. Instructors in man-dominated

disciplines who do not make students feel as if they have learned from their course are much more likely to receive lower overall student evaluation scores than those who do make their students feel as if they have learned.

For perceived role-neutral faculty, three variables were determined to be weighted differently for men versus women instructors. For men, Content-Related-Assignments on the latent concept Overall and Instructor-Organized on the latent concept Instructor mattered more for men than they did for women. For women, Content-Thought-Provoking on the latent concept Overall mattered more than for men. Content-Thought-Provoking also mattered more in the measurement for women than for men among perceived role congruent instructors but not for role incongruent instructors. This indicates that for women who are perceived role congruent or neutral, having thought-provoking content in the course has a greater effect on their SEI scores than for role congruent or neutral men. Thus, women in women-dominated or neutral disciplines should thoughtfully consider the content of their courses in order to improve their student evaluation scores. This could be the case because women are perceived to be experts in women-dominated disciplines and, to an extent, in neutral disciplines thus there is a higher bar set, and/or students are more critical of the material they include in their classes than their men and perceived role incongruent women peers.

For perceived role incongruent and role neutral faculty, once the models were constrained based on the determined measurement invariance there were not statistically significant differences in the means of either latent construct, Overall and Instructor. However, for perceived role congruent instructors, statistically significant differences in the means of the latent constructs persisted even when the model was properly constrained. This indicates that student evaluation forms may be more sensitive to differences in student perceptions of their

instructors when instructors are teaching in perceived role congruent disciplines than when they are teaching in perceived role incongruent or role neutral disciplines. This may mean that when instructors are teaching in disciplines for which they are perceived to be an expert based on their gender presentation, students are more critical of the quality of the course and instructor. Perceived role incongruent instructors may not have as high of a bar set for them by students because of their perceived "lack of fit" with the discipline, a benefit that also seeps into role-neutral disciplines. Students may see a woman teaching an engineering course and expect the class to not be as good as if a man were teaching it thus potentially leading to higher evaluation scores for the woman when she does teach well. Women who teach poorly in engineering may also receive more positive evaluations than men in engineering who teach poorly because she is not expected to teach well whereas he is expected to do so. This phenomenon may also lead to lower evaluation scores for a man teaching an engineering course even if his teaching is average or above because he is expected to be an expert in the subject and therefore teach it well so when he does, he does not receive any special benefit for doing so. The opposite pattern of positive and negative effects would occur for men and women teaching in a woman-dominated field such as English—women who teach well will not receive any special benefit for doing so but will be especially penalized when they teach poorly whereas men will receive higher scores regardless of their actual teaching. Further research should continue to determine the causal mechanism behind the measurement invariance of perceived role congruent instructors.

On the whole, these mixed results indicate that it is incredibly important to consider both the gender and perceived role (in)congruity of the instructor when completing measurement invariance testing and comparing the mean student evaluation scores of instructors. Utilizing the proper models with constrained observed variables, the results of these analyses indicate that the

perceived gender role congruity between the instructor and the discipline in which they are teaching can affect how students evaluate their course. For perceived role congruent instructors, women receive slightly higher scores than their men counterparts on both latent concepts (Overall and Instructor) while there are no statistically significant differences between the scores of men and women who teach in role incongruent or role neutral disciplines. These results indicate that gender matters, but how it matters depends upon the perceived gender role congruity of the discipline and instructor. Women instructors who teach in women-dominated fields are likely to receive slightly higher scores than men who teach in men-dominated disciplines. Women who teach in man-dominated disciplines or role neutral disciplines are likely to receive scores that are similar to their respective men who teach in feminine or role neutral disciplines.

These results stand in contrast to the vast literature that has found that women instructors tend to score worse on their student evaluations than men (Basow 1995; Boring et al. 2016; El-Alayli et al. 2018; MacNell et al. 2015b). The results of this study indicate that when models are properly constrained to account for measurement differences between faculty of different genders, there are very limited differences in student evaluation scores between men and women instructors. In fact, when the models are properly constrained perceived role congruent women instructors may be at an advantage as compared to their perceived role congruent men peers.

There is evidence that suggests that student evaluation scores tend to be higher in women-dominated disciplines such as the humanities than in men-dominated disciplines such as STEM fields (Basow and Montgomery 2005; Wachtel 1998). However, the results of this study indicate that while perceived role congruent women score statistically significantly higher than role congruent men, there were not statistically significant differences between the scores of

perceived role incongruent women and men. This means that while men in women-dominated disciplines do not score higher than women in men-dominated disciplines, women in women-dominated disciplines do score higher than men in men-dominated disciplines. Therefore, even when women teach in men-dominated disciplines in which student evaluation scores are typically lower, their scores are actually on par with their men colleagues who are teaching in women-dominated disciplines that typically receive higher student evaluation scores. Thus, when models are properly constrained, there may not be as prominent of differences in the student evaluation scores between men and women instructors especially when the perceived role (in)congruity of the instructor is considered. Additionally, the differences that do persist may be the opposite of previous findings with role congruent women scores slightly higher than role congruent men on both the Overall and Instructor latent constructs.

Limitations

Though this study adds significantly to the student evaluation and role congruities literatures, this study is not without its own limitations. The data were limited by what was available through the institution. There were only a finite number of student evaluation questions asked systematically enough for analyses to be completed. The finite number of systematically asked student evaluation questions limited the statistical models that could be conducted. For more detailed analyses, more student evaluation questions would need to be asked systematically across all students, courses, and instructors.

Furthermore, the data tends to skew relatively high with the mean scores of the seven variables ranging from 3.897 to 4.419 on five-point scales. These mean scores indicate that students are generally positive in their quantitative evaluation of their instructors. The relative homogeneity of the data limits the amount of differences that can be found in the data. The

institution may want think about what the goal of student evaluations of instruction are because if the goal is to determine the best and worst instructors the current forms are not leading to substantial differences between instructors.

There are also many other factors which were not included in these analyses that have been shown to affect student evaluation scores such as students' anticipated grades, student gender, course level, course type (mandatory versus elective), and the number of credits a course is worth (Kalender 2015). Additionally, research has found that women instructors tend to be tasked with more student support work outside of the classroom such as advising, mentoring, and providing feedback on work (El-Alayli et al. 2018). In the experiment conducted by MacNell, Drescoll, and Hunt (2015b), instructors who were perceived to be male were rated higher on six interpersonal measures than the perceived female instructor even with all communication and grading procedures held equal. These findings indicate that female instructors may be expected to be more interpersonal than male instructors. Male instructors, on the other hand, are not held to this interpersonal expectation and are therefore rewarded as having gone "above and beyond" when they do display interpersonal traits (MacNell et al. 2015b).

Though the results of the study conducted in this chapter indicate that there may not be extreme gender-based differences in student evaluation scores and that perceived role congruent women may actually receive higher scores than perceived role congruent men, these women may be putting in additional labor for students outside of teaching and "going above and beyond" in other ways in order to receive these higher scores. Thus, though women appear to have an advantage in this study, they may be working even harder than their men colleagues to receive the same or slightly higher evaluation scores. A study of the instructor time and the amount of labor being done for students outside of the classroom would be useful to determine why women

instructors in this context received scores that are on par with or even slightly higher than men instructors.

Future Studies

The results of this study indicate that future student evaluation studies should test their models for measurement invariance *before* comparing the means of different faculty groups. By determining the best fitting model, more accurate between-group comparisons can be made. Furthermore, future studies should use these better-fitting models to compare the means between other groups of faculty beyond women and men. For example, future studies could examine the differences between faculty of different races/ethnicities, ranks, ages, etc. Additionally, future studies should further tease out how perceptions of instructor role congruity affects evaluations in combination with these other instructor characteristics. This study illustrated that the perceived role congruity between instructor gender and discipline affects whether or not there are differences in SEI scores. Future studies should examine the ways in which perceived gender role congruity in combination with other faculty characteristics may affect SEI scores such as faculty race/ethnicity, age, position, and teaching style.

CONCLUSION

The results of this study indicate that when examining how target identities affect subjective evaluations, attention needs to be given to both potential measurement invariance and perceptions of target role (in)congruity. Measurement invariance testing is a crucial step that needs to be conducted before comparing the means between two groups. Without measurement invariance testing and properly constraining the models as necessary, the means of the groups cannot be meaningfully compared because they are not even being measured in the same way. Thus, every study and evaluation form should utilize either pre-validated measures which have

been tested and adjusted for measurement invariance or the researchers should go through the steps of testing and constraining for measurement invariance prior to comparing the means between the groups. With respect to perceptions of target role (in)congruity, it is clear from these results that the effects of gender alone do not highlight the nuanced ways in which students' completion of subjective evaluations of instructors is affected by the identities of the instructors, in this case, perceived role (in)congruity. These results call into question what other faculty identities might affect students' SEI responses. One possible target identity that may affect student evaluations of instruction is the race/ethnicity of the instructor. In the next study, I examine the ways in which students' evaluations of instruction are affected by their perceptions of the gender, discipline, *and* race/ethnicity of their instructors.

CHAPTER 3: QUANTITATIVE ANALYSES OF STUDENT EVALUATIONS OF INSTRUCTION WITH ATTENTION TO FACULTY GENDER, RACE/ETHNICITY, AND ROLE (IN)CONGRUITY

INTRODUCTION

Subjective evaluations are assessments that are highly affected by the characteristics of the target of the evaluation such as their gender, race, and age (Arbuckle and Williams 2003; Smith et al. 2001). Student evaluations of instruction (SEIs) are one type of assessment that are utilized in almost every institution of higher education to evaluate instructor performance. Previous research has shown that SEIs are highly affected by the status characteristics of the instructor being evaluated (Boring et al. 2016; Smith et al. 2001).

In the previous study, it was shown that students' evaluations of their instructors are affected by the gender and the gender-dominance of the discipline in which they teach. While the findings of the previous study add much to the literature on subjective evaluations and specifically student evaluations of instruction, there are many other instructor characteristics that may affect students' evaluations of their instructors. One such characteristic is the race of the instructor being evaluated (Anderson and Smith 2005; Reid 2010; Smith and Hawkins 2011). While instructor race has been shown to affect student perceptions of their instructors, there is a dearth of research that examines how the race, gender, and gender-dominance of the discipline may interact to affect these perceptions.

In this study, student evaluations of instruction are quantitatively analyzed with consideration of the gender, race/ethnicity, and perceived gender role (in)congruity of the course instructor. Through these analyses, I seek to answer the question are students' subjective evaluations of their instructors affected by the race/ethnicity and perceived gender role (in)congruity of the instructor? In the next section theories of gender role congruity, specifically

Role Congruity Theory (RCT) and Status Incongruity Hypothesis (SIH), are described as well as some of the previous research on student evaluations of instruction (SEIs) with particular focus on the effects of instructor race/ethnicity.

LITERATURE REVIEW

The root social psychological theories used in this study and throughout this dissertation are Role Theory and Social Role Theory. Role Theory considers the roles people occupy and the behavioral, attitudinal, and value expectations associated with those roles (Jacobs 2018b). People occupy multiple roles at multiple levels and they must learn to navigate the potentially conflicting expectations associated with those roles (Jacobs 2018b; Lynch 2007). Roles are diffuse, they occur in most situations such as gender, race, and age, or specific, they only occur in specific circumstances such as occupations or parental status (Diekman and Schneider 2010b; Koenig and Eagly 2014). The demands of a specific role such as occupation can affect how much a person's behaviors, values, and beliefs are determined by their diffuse roles and vice versa (Eagly et al. 2000). A person's many diffuse and specific roles may carry conflicting role expectations.

Social Role Theory focuses more on the internal and external effects of occupying multiple roles and especially multiple roles with conflicting expectations (Eagly and Karau 2002). *Social roles* refer to the shared expectations associated with people who are members of a particular social category such as gender or race (Eagly et al. 2000). Overrepresentation of a particular social group in a social role leads to the development of stereotypes associated with people from that social group with the social role such as women being associated with childcare thus women being assumed to be warm and caring (Koenig and Eagly 2014). For a further explanation of these two theories, see Chapters 1 and 2.

Theories of Congruity

Role Congruity Theory (RCT) and Status Incongruity Hypothesis (SIH) are further derivations of Role Theory and Social Role Theory that consider how a persons' multiple statuses interact to affect perceptions and evaluations of them. Specifically, theories of gender role congruity posit that people who occupy specific roles which are incongruent with their gender will receive sanctions such as negative evaluations and other punishments (Diekman and Eagly 2008). Many researchers have studied the effects of gender role congruity in leadership contexts and have found that the gender role congruity of a leadership candidate affects evaluations of the potential success as a leader (Garcia-Retamero and López-Zafra 2006). Specifically, people tend to evaluate women as less capable leaders than their man counterparts (Garcia-Retamero and López-Zafra 2006). Additionally, respondents attribute failures to something internal to the woman leader and attribute any successes women leaders experience to external factors (Garcia-Retamero and López-Zafra 2006). On the other hand, respondents attribute failures of men to external causes and attribute successes of men to internal causes (Garcia-Retamero and López-Zafra 2006). These results highlight the numerous ways in which the perceived gender role congruity of a leader can affect evaluations of the leader, their failures, and their successes.

One major limitation of theories of gender role congruity is the lack of inclusion of diffuse statuses beyond gender. There are many diffuse roles that are present and affect a person in most situations other than gender (Koenig and Eagly 2014). Salient social statuses such as race, age, and class operate in ways that are similar to gender in that they "involve(s) cultural beliefs and distributions of resources at the macro level, patterns of behavior and organizational practices at the interactional level, and selves and statuses at the individual level" (Ridgeway and

Correll 2004:510–11; Ridgeway and Smith-Lovin 1999). Thus, there is reason to believe that other salient social statuses such as race, age, and class may operate in similar ways to gender and thus moderate considerations of gender role (in)congruity (Andreoletti, Leszczynski, and Disch 2015; Ridgeway and Correll 2004). Race, in particular, may be a diffuse role that affects perceptions and may even moderate the effects of considerations of gender role (in)congruity. As described in detail in Chapter 1, racially/ethnically minoritized persons, much like women, have historically been discriminated against in workplace and leadership contexts in the United States. The historical exclusion of racially/ethnically minoritized persons from leadership roles may lead to different perceptions and evaluations of racially/ethnically minoritized persons in leadership contexts much like there are different perceptions and evaluations of women in leadership contexts as compared to men.

For example, Livingston, Rosette, and Washington (2012) found that when examining gender role congruity effects in leadership contexts, the race of the leader being evaluated moderates perceptions and evaluations of them. Specifically, Black women leaders were evaluated as positively as White men leaders and more positively than both Black men leaders and White women leaders (Livingston et al. 2012). Biernat and Seko (2013) found similar results in their two-part study in which they compared evaluations of the members of hypothetical dyads of coworkers with varied racial and gender identities. White men tended to be evaluated as more competent than White women when they were paired together but when White women and Black men were paired together there were no differences in the competency evaluations (Biernat and Sesko 2013). Furthermore, the evaluations of competence of Black women were not significantly different from their partner's when they were paired with White men or Black men (Biernat and Sesko 2013). The results of these two studies indicate that perceptions of gender role congruity

effects, while powerful on their own, may also be moderated by a target's other diffuse roles such as their race. Thus, it is critical that gender role congruity researchers consider more salient social roles in the examination of perceived gender role (in)congruity.

Study Context: Student Evaluations of Instruction

One such area in which both race and gender have been shown to affect evaluations is in the study of subjective evaluations and, more specifically, the study of student evaluations of instruction (SEIs) in higher education (Basow 1995; Basow, Phelan, and Capotosto 2006; Bavishi et al. 2010). Research on SEIs has consistently shown that women tend to receive lower SEI scores than men (Basow 1995; Boring et al. 2016; El-Alayli et al. 2018; MacNell et al. 2015a). Not only are SEIs biased against women instructors but they are biased enough to cause more effective women instructors to receive lower SEI scores than less effective men instructors (Boring et al. 2016).

Furthermore, the race of an instructor can also affect student evaluation scores with White instructors more likely to receive higher scores than racially/ethnically minoritized instructors (Reid 2010; Smith and Hawkins 2011). On the popular RateMyProfessors.com instructor reviewing website, the best-ranked professors are more likely to be White while the worst-ranked professors are more likely to be Black or Asian (Reid 2010). On student evaluations, Black faculty mean evaluation scores are lower than White and other racial groups across a wide spectrum of measures (Smith and Hawkins 2011). Experimental research has also found similar results with White professors tending to receive higher ratings of favorability and trust than Black professors (Aruguete, Slater, and Mwaikinda 2017).Quasi-experimental research has found similar results with women and racially/ethnically minoritized instructors tending to

receive statistically significantly lower overall student evaluation scores than their men and White counterparts (Chávez and Mitchell 2020).

Even when all other factors such as students' final grades are held equal, the gender and racial disparities in student evaluation scores persist (Chávez and Mitchell 2020). Gender and race have also been shown to have an interaction effect on student evaluations (Anderson and Smith 2005; Chávez and Mitchell 2020). In an experiment in which respondents evaluated hypothetical instructors based on their course syllabi which varied by teaching style, gender, and race/ethnicity, multiple interaction effects were found (Anderson and Smith 2005). White women with a strict teaching style were viewed as warmer than Latinx women professors with a strict teaching style while Latinx women professors with a lenient teaching style were viewed as warmer than White women with a lenient teaching style (Anderson and Smith 2005).

Research by Bavishi, Madera, and Hebl (2010) added another dimension to the study of the effects of instructor race on students' evaluations of them by considering the discipline of the instructor. Their results indicate that while White instructors, in general, tend to be rated higher in competency and legitimacy than Black or Asian instructors, science instructors regardless of race tended to be ranked as more competent and more legitimate than humanities instructors (Bavishi et al. 2010). Additionally, there was a significant interaction between instructor race and legitimacy but not competency. Black professors in the humanities were perceived as less legitimate than White professors in science (Bavishi et al. 2010). Thus, students' evaluations of their instructors may be affected by the gender (e.g. MacNell et al. 2015), race (e.g. Smith and Hawkins 2011), and discipline (e.g. Bavishi et al. 2010) of the instructor in question. Furthermore, the results of Study 1 (Chapter 2) indicate that instructor gender and their perceived gender role (in)congruity with the discipline in which they teach combine to affect student

evaluations. However, there is a lack of research that examines the ways in which instructor gender, race/ethnicity, and perceived role (in)congruity may combine to affect students' evaluations. It is therefore critical that researchers take at least these three intersecting identities of instructors into account when studying student evaluations of instruction.

Additionally, it is critical, as outlined in Study 1 (Chapter 2), that studies of student evaluations of instruction begin with tests for measurement invariance prior to conducting any mean comparisons. Without accounting for any measurement invariance that may be occurring, any results that are presented may be biased due to measurement differences due to the gender, race, and/or discipline of the instructors being evaluated. In this study, not only are more intersectional identities of instructors taken into consideration through the inclusion of instructors' gender and racial/ethnic identities as well as the gender-dominance in which they teach but also measurement invariance is tested and accounted for before any mean comparisons are completed. These are two unique contributions this study makes to research on student evaluations, studies of role congruity, and studies of subjective evaluations more generally.

DATA

To complete this research, student evaluations of instruction data as well as human resources data from a large research-intensive land-grant university in the Appalachian region of the United States were utilized. See Chapter 1 for a complete description of the data cleaning, merging, and variable creation process. The seven quantitative questions utilized in Study 1 (Chapter 2) are used in this study and presented in Table 13. See Study 1 (Chapter 2) for a full description of the seven questions included in these analyses.

Table 13: Variables from Student Evaluations of Instruction

Variable	Variable Name	Student Evaluation Question
Content Related to Assignments	Content-Related-Assignments	Course content was related to graded assignments
Content Thought Provoking	Content-Thought-Provoking	Course content was thought provoking
Material Being Useful	Material-Useful	The course materials were useful to course objectives
Positive Learning Environment	Positive-Learning-Environment	The instructor fostered a positive learning environment
Instructor Organization	Instructor-Organized	The instructor was well organized
Instructor Feedback	Instructor-Feedback	The instructor provided helpful feedback
Overall Learning	Overall-Learning	Overall my learning in this course was

METHODS AND ANALYSES

The finalized dataset was analyzed using STATA statistical software. A full description of the analytical process can be found in Study 1 (Chapter 2). In summary, statistical analyses started with an Exploratory Factor Analysis of all of the data together. Analyses were then divided by faculty congruity group, those being role congruent, role incongruent, and role neutral. The structure of the models were confirmed through Confirmatory Factor Analyses (CFA). MIMIC models were the next step in Study 1 (Chapter 2), however, they were not completed in this study as they require binary exogenous covariates whereas this study utilized four groups those being White men, racially/ethnically minoritized men, White women, and racially/ethnically minoritized women. Thus, the next step was to conduct grouped Structural Equation Models (SEM). The SEMs were grouped on instructor gender and race/ethnicity. The grouped SEMs were completed to determine if there is measurement invariance based on faculty race/ethnicity, gender, and gender role congruity. Post-estimation testing further examined which variables exhibited measurement invariance in each of the models.

RESULTS

The datasets included 52,383 responses for faculty perceived to be role congruent, 51,684 responses for faculty perceived to be role incongruent, and 25,076 responses for faculty perceived to be role neutral. The analyses conducted utilized listwise deletion, so any incomplete responses were automatically dropped from the analyses. For example, if an evaluation included responses for all but one of the questions included in the models, that evaluation was not included in the analyses. Table 14 describes the breakdown of the sample by faculty group, instructor gender, and instructor race/ethnicity. A "response" is a single student evaluation from one student for one professor about one course.

Table 14: SEI Response Counts by Instructor Gender, Race/Ethnicity, and Gender Role Congruity Group

Faculty Group	White Men	White Women	Racially/ethnically Minoritized Men	Racially/ethnically Minoritized Women	Total
Gender Role Congruent	23,870	17,256	7,925	3,332	52,383
Gender Role Incongruent	11,108	9,952	2,364	2,674	26,098
Gender Role Neutral	4,889	6,018	944	791	12,642

See Study 1 (Chapter 2) for the results for the Exploratory Factor Analysis and Confirmatory Factor Analyses as there were no changes made to these models in these analyses. There are no changes to the results of the EFA and CFAs because those models were completed on all of the data and all of the data within each role congruity group, respectively. Thus, the results are unaffected by the grouping that occurs in the subsequent testing meaning that the results from Study 1 (Chapter 2) are the same results as would be presented here. Following the format of Study 1 (Chapter 2), the results of the grouped SEM models will be presented in each

of the instructor groups based on perceived level of congruity: role congruent, role incongruent, and role neutral.

Gender Role Congruent Faculty

Grouped Structural Equation Models

A grouped SEM was conducted utilizing the CFA model from Study 1 (Chapter 2) and grouped based on instructor gender and race/ethnicity: gender role congruent White men, gender role congruent White women, gender role congruent racially/ethnically minoritized men, and gender role congruent racially/ethnically minoritized women. To test for configural invariance, the means of the latent concepts were set equal to zero but there were no constraints placed on the groups. The results of the same form equivalence model which tests for configural invariance can be found in Table 15. The loadings were all substantial and statistically significant and the model fit statistics all indicated that the model fit well ($\chi 2(48)=1148.62$, $p \leq 0.001$; CFI=0.996, TLI=0.992, SRMR=0.011, RMSEA=0.042).

Table 15: Gender Role Congruent Same Form Equivalence Model

	Same Form Equivalence							
	White Men (*N*=23,870)		White Women (*N*=17,256)		Racially/ethnically Minoritized Men (*N*=7,925)		Racially/ethnically Minoritized Women (*N*=3,332)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	a	0.84***	a	0.85***	a	0.85***	a	0.87***
Content-Thought-Provoking	1.05***	0.80***	1.11***	0.83***	1.03***	0.80***	1.00***	0.82***
Material-Useful	1.16***	0.90***	1.19***	0.90***	1.13***	0.90***	1.16***	0.92***
Overall-Learning	0.49***	0.32***	0.41***	0.27***	0.43***	0.28***	0.30***	0.21***
Instructor								
Positive-Learning-Environment	1.41***	0.85***	1.29***	0.85***	1.29***	0.86***	1.08***	0.85***
Instructor-Organized	1.44***	0.82***	1.39***	0.82***	1.29***	0.82***	1.22***	0.86***
Instructor-Feedback	1.67***	0.88***	1.58***	0.87***	1.51***	0.88***	1.36***	0.90***
Overall-Learning	a	0.50***	a	0.49***	a	0.56***	a	0.62***
Mean Overall	a	a	a	a	a	a	a	a
Mean Instructor	a	a	a	a	a	a	a	a
R^2	0.979		0.981		0.980		0.984	
$\chi 2$	*df*=48, 1148.62***							
CFI	0.996							
RMSEA	0.042							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*p<0.05; **p<0.01; ***$p\leq$0.001

To test for metric invariance, the means of the latent concepts were still set to equal zero and the measurement coefficients were constrained to be equal across the groups. The loadings were once again all substantial and statistically significant and the model fit statistics all indicated that the model fit well ($\chi 2(66)=1343.55$, $p\leq0.001$; CFI=0.995, TLI=0.994, SRMR=0.020, RMSEA=0.038). The results of this model are presented in Table 16.

Table 16: Gender Role Congruent Equal Loadings Model

	Equal Loadings							
	White Men (*N*=23,870)		White Women (*N*=17,256)		Racially/ethnically Minoritized Men (*N*=7,925)		Racially/ethnically Minoritized Women (*N*=3,332)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	a	0.84***	a	0.86***	a	0.84***	a	0.86***
Content-Thought-Provoking	1.06***	0.80***	1.06***	0.82***	1.06***	0.80***	1.06***	0.83***
Material-Useful	1.17***	0.90***	1.17***	0.90***	1.17***	0.90***	1.17***	0.91***
Overall-Learning	0.43***	0.28***	0.43***	0.29***	0.43***	0.29***	0.43***	0.30***
Instructor								
Positive-Learning-Environment	1.31***	0.85***	1.31***	0.85***	1.31***	0.85***	1.31***	0.86***
Instructor-Organized	1.37***	0.82***	1.37***	0.81***	1.37***	0.83***	1.37***	0.85***
Instructor-Feedback	1.58***	0.88***	1.58***	0.87***	1.58***	0.88***	1.58***	0.90***
Overall-Learning	a	0.53***	a	0.48***	a	0.54***	a	0.53***
Mean Overall	a	a	a	a	a	a	a	a
Mean Instructor	a	a	a	a	a	a	a	a
R^2	0.978		0.981		0.980		0.983	
χ2	*df*=66, 1343.55***							
CFI	0.995							
RMSEA	0.038							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

A likelihood-ratio test was then used to compare the two models. The likelihood ratio test indicated that the invariant loading model performs statistically significantly worse than the same form model ($\chi 2(18)=194.94$, $p\leq0.001$). Thus, we should not constrain the loadings to be equal. This means that there are statistically significant differences between at least two of the four instructor groups in the meaning of the latent variables Instructor and Overall when measured

with the observed variables used in these models. Because the overall model fit of the metric invariance model is worse than the configural invariance model this means that at least one loading is not equivalent across the groups. Metric invariance is therefore *not* supported.

A postestimation test indicates that all but one of the observed variables in the model differ significantly between at least two of the gender role congruent instructor groups in the level of importance they carry in their measurement of the latent concepts. The output shows significant chi-squared values for all of the variables in the model except for Instructor-Feedback (5.784, p=0.1226) which can be seen in Table 17. This means that all of the variables except for Instructor Feedback (Instructor-Feedback) differ significantly on their levels of importance between at least two of the instructor groups those being White men, White women, racially/ethnically minoritized men, and racially/ethnically minoritized women.

Table 17: Gender Role Congruent Test of Group Invariance of Parameters

Latent Variables	Observed Variables	Score Test		
		$\chi 2$	*df*	*p*-value
Overall	Content-Related-Assignments	36.190	1	<0.001
	Content-Thought-Provoking	42.140	1	<0.001
	Material-Useful	13.231	1	0.0042
	Overall-Learning	33.502	1	<0.001
Instructor	Positive-Learning-Environment	50.642	1	<0.001
	Instructor-Organized	45.621	1	<0.001
	Instructor-Feedback	5.784	1	0.1226
	Overall-Learning	20.592	1	0.0001

A partial invariant loadings model can be run to compare the mean scores between the instructor groups. In the partial invariant loadings model, the loadings for the one variable which was determined to not be measured differently was constrained to be equal while the loadings for the seven variables that were determined to be different were allowed to vary. Since all of the variables measuring the latent concept Overall were determined to be measured differently, one

of the variables needs to be constrained to run the partial invariant model. To test for differences in the constrained variable, the constrained variable will be rotated in order to test for differences in every variable. The loadings for the partially invariant model were all positive but not statistically significant. The model fit statistics all indicated that the model fit well regardless of which variable on Overall was constrained (Constrained Content-Related-Assignments $\chi 2(62)=1326.64$, $p\leq 0.001$; CFI=0.995, TLI=0.993, SRMR=0.012, RMSEA=0.039; Constrained Content-Thought-Provoking $\chi 2(62)=1326.64$, $p\leq 0.001$; CFI=0.995, TLI=0.993, SRMR=0.012, RMSEA=0.039). The results of the partial invariant loadings model can be found in Tables 18 and 19.

According to the partial invariant loadings model, Content Related to Assignments (Content-Related-Assignments) on the latent concept Overall carries the least weight for perceived gender role congruent White women (0.90) followed by White men and racially/ethnically minoritized men (0.96), and the most weight for racially/ethnically minoritized women (1.00). The opposite pattern emerged for the variable Content Thought Provoking (Content-Thought-Provoking) on the latent concept Overall which carries the least weight for racially/ethnically minoritized women (1.00) followed by racially/ethnically minoritized men (1.04), White men (1.05), and the most eight for White women (1.11).

The weights for the variable Material-Useful on the latent concept Overall were more susceptible to variation based on which other variable was constrained in the model. When Content-Related-Assignments was constrained, Material-Useful carries the least weight for racially/ethnically minoritized men (1.13), followed by White men (1.16)[8], racially/ethnically

[8] All tables were rounded to two decimal places however, the full reported value was considered when selecting which was higher. In this case, the coefficient for White men was 1.158 and for racially/ethnically minoritized women it was 1.16.

minoritized women (1.16), and the most weight for White women (1.19). WhenContent-Thought-Provoking was constrained, Material-Useful carries the least weight for White women (1.19) followed by racially/ethnically minoritized men (1.09), White men (1.11), and the most weight for racially/ethnically minoritized women (1.16).

Overall-Learning on the latent concept Overall carries the least weight for perceived gender role congruent racially/ethnically minoritized women (0.31) followed by White women (0.40), racially/ethnically minoritized men (0.43), and the most weight for White men (0.49). On the latent concept Instructor, Overall-Learning carries the least weight for perceived gender role congruent White men (0.56) followed by White women (0.61), racially/ethnically minoritized men (0.62), and the most weight for racially/ethnically minoritized women (0.68). The most and least influenced groups switched for the variable Overall-Learning depending on which latent concept it was measuring.

The variable Positive-Learning-Environment on the latent concept Instructor carries the least weight for perceived gender role congruent racially/ethnically minoritized women (0.74) followed by White women (0.77), White men (0.79), and the most weight for racially/ethnically minoritized men (0.80).[9] Instructor-Organized on the latent concept Instructor carries the least weight for racially/ethnically minoritized men (0.80) followed by White men (0.80), White women (0.82), and the most weight for racially/ethnically minoritized women (0.84).[10]

With respect to the means, there are significant differences between the means of the different instructor groups for the latent construct Overall. The differences in the means for the Overall latent concept vary depending on which instructor group is the reference group. The

[9] When Content-Thought-Provoking was constrained, the pattern of the results was the same when considering the full coefficients rather than the rounded values.
[10] Once again the full value was considered and the pattern of results between the model with Content-Related-Assignments and Content-Thought-Provoking remained the same.

results in Tables 18 and 19 use White men as the reference group. The remaining reference group rotations can be found in Appendices 1-6. There are some general patterns to the results such that the means for gender role congruent White women are statistically significantly higher than of any other instructor groups. White men and racially/ethnically minoritized men tend to have the least amount of differences from the reference group. The mean score for White men is statistically significantly lower than that of White women but higher than that of racially/ethnically minoritized men and racially/ethnically minoritized women. The mean score for racially/ethnically minoritized men is lower than that of White men and White women but higher than that of racially/ethnically minoritized women. The mean scores for racially/ethnically minoritized women are consistently lower than the means of the reference groups. However, the intensity of the difference between of the mean scores of gender role congruent racially/ethnically minoritized women and the reference groups varies the most of any instructor group with the mean score of racially/ethnically minoritized women having the least difference from the mean of racially/ethnically minoritized men, the most difference from the mean of White women, and falling in the middle of the mean differences when White men are the reference group. In sum, the means of women and especially White women tend to be most different from the reference group on the latent concept Overall with the mean scores of perceived gender role congruent White women are higher than the reference group while the mean scores for racially/ethnically minoritized women are lower than the reference group. The means of perceived gender role congruent racially/ethnically minoritized men are lower than all groups except for racially/ethnically minoritized women and the means for perceived gender role congruent White men are higher than all groups except White women.

The results further indicate that there are *not* statistically significant differences between the scores of perceived gender role congruent White men and the other three instructor groups[11]—White women (0.29, p=0.97), racially/ethnically minoritized men (-0.11, p=0.973), and racially/ethnically minoritized women (-0.07, p=0.973)—on the latent construct Instructor. These results indicate that when the model is properly constrained for measurement differences, there are not score differences on the observed variables for the latent construct Instructor based on instructor gender, race/ethnicity, and gender role congruity.

[11] The means comparison models were run with each of the four instructor groups as the reference group. The results of the other three rotations can be found in Appendices 1-6. There was one significant difference between the means of racially/ethnically minoritized women and racially/ethnically minoritized men, however this difference depends on which variable is constrained on the latent concept Overall and appears to be practically insignificant. Since all of the models indicate the same results except for this one small difference, only one is discussed in text.

Table 18: Gender Role Congruent Partial Invariant Model Means Comparison

	Content-Related-Assignments Constrained							
	White Men (*N*=23,870)		White Women (*N*=17,256)		Racially/ethnically Minoritized Men (*N*=7,925)		Racially/ethnically Minoritized Women (*N*=3,332)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	a	0.84***	a	0.85***	a	0.85***	a	0.87***
Content-Thought-Provoking	1.05***	0.80***	1.11***	0.83***	1.04***	0.80***	1.00***	0.82***
Material-Useful	1.16***	0.90***	1.19***	0.90***	1.13***	0.89***	1.16***	0.92***
Overall-Learning	0.49***	0.32***	0.40***	0.27***	0.43***	0.28***	0.31***	0.22***
Instructor								
Positive-Learning-Environment	0.79	0.85***	0.77	0.85***	0.80	0.86***	0.74	0.85***
Instructor-Organized	0.81	0.82***	0.82	0.82***	0.80	0.82***	0.84	0.86***
Instructor-Feedback	0.94	0.88***	0.94	0.87***	0.94	0.88***	0.94	0.90***
Overall-Learning	0.56	0.50***	0.61	0.50***	0.62	0.56***	0.68	0.61***
Mean Overall	a	a	0.11***	0.16***	-0.06***	-0.07***	-0.10***	-0.12***
Mean Instructor	a	a	0.29	0.33***	-0.11	-0.10***	-0.07	-0.07***
R^2	0.978		0.981		0.980		0.984	
χ2	*df*=62, 1326.64***							
CFI	0.995							
RMSEA	0.039							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Table 19: Gender Role Congruent Partial Invariant Model Means Testing Rotating Constraints

	Positive-Learning-Environment Constrained							
	White Men (*N*=23,870)		White Women (*N*=17,256)		Racially/ethnically Minoritized Men (*N*=7,925)		Racially/ethnically Minoritized Women (*N*=3,332)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.96***	0.84***	0.90***	0.85***	0.96***	0.85***	1.00***	0.87***
Content-Thought-Provoking	a	0.80***	a	0.83***	a	0.80***	a	0.82***
Material-Useful	1.11***	0.90***	1.07***	0.90***	1.09***	0.89***	1.16***	0.92***
Overall-Learning	0.47***	0.32***	0.36***	0.27***	0.41***	0.28***	0.31***	0.22***
Instructor								
Positive-Learning-Environment	0.79	0.85***	0.76	0.85***	0.79	0.86***	0.74	0.85***
Instructor-Organized	0.81	0.82***	0.81	0.82***	0.79	0.82***	0.84	0.86***
Instructor-Feedback	0.93	0.88***	0.93	0.87***	0.93	0.88***	0.93	0.90***
Overall-Learning	0.56	0.50***	0.60	0.50***	0.62	0.56***	0.68	0.61***
Mean Overall	a	a	0.13***	0.16***	-0.06***	-0.07***	-0.10***	-0.12***
Mean Instructor	a	a	0.29	0.33***	-0.11	-0.10***	-0.07	-0.07***
R^2	0.978		0.981		0.980		0.984	
χ^2	*df*=62, 1326.64***							
CFI	0.995							
RMSEA	0.039							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*p<0.05; **p<0.01; ***p≤0.001

Gender Role Incongruent Faculty

Grouped Structural Equation Models

A grouped SEM was conducted utilizing the CFA model from Study 1 (Chapter 2) and grouped based on instructor gender and race/ethnicity. To test for configural invariance, a same form equivalence model was run in which the means of the latent concepts were set to equal zero

but there were no constraints placed on the groups. The full results of the same form equivalence model for perceived gender role incongruent faculty can be found in Table 20. The loadings were all substantial and statistically significant and the model fit statistics all indicated that the model fit well ($\chi 2(48)=599.98$, $p \leq 0.001$; CFI=0.996, TLI=0.992, SRMR=0.014, RMSEA=0.042).

Table 20: Gender Role Incongruent Same Form Equivalence Model

	Same Form Equivalence							
	White Men (*N*=11,108)		White Women (*N*=9,952)		Racially/ethnically Minoritized Men (*N*=2,364)		Racially/ethnically Minoritized Women (*N*=2,674)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	a	0.86***	a	0.85***	a	0.84***	a	0.83***
Content-Thought-Provoking	1.06***	0.83***	1.03***	0.79***	1.05***	0.80***	1.17***	0.83***
Material-Useful	1.14***	0.90***	1.15***	0.90***	1.15***	0.89***	1.22***	0.89***
Overall-Learning	0.36***	0.25***	0.48***	0.32***	0.38***	0.25***	0.57***	0.35***
Instructor								
Positive-Learning-Environment	1.28***	0.85***	1.40***	0.87***	1.29***	0.84***	1.53***	0.87***
Instructor-Organized	1.31***	0.82***	1.46***	0.84***	1.30***	0.82***	1.57***	0.83***
Instructor-Feedback	1.51***	0.87***	1.68***	0.89***	1.55***	0.84***	1.79***	0.89***
Overall-Learning	a	0.53***	a	0.48***	a	0.51***	a	0.47***
Mean Overall	a	a	a	a	a	a	a	a
Mean Instructor	a	a	a	a	a	a	a	a
R^2	0.981		0.981		0.977		0.982	
$\chi 2$	*df*=48, 599.98***							
CFI	0.996							
RMSEA	0.042							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p \leq 0.001$

To test for metric invariance, an equal loadings model was run in which the means of the latent concepts were still set to equal zero and the measurement coefficients were constrained to

be equal across the groups. The loadings were once again all substantial and statistically significant and the model fit statistics all indicated that the model fit well ($\chi 2(66)=672.80$, $p \leq 0.001$; CFI=0.995, TLI=0.994, SRMR=0.022, RMSEA=0.038). The full results of the equal loadings model for perceived gender role incongruent faculty can be found in Table 21.

Table 21: Gender Role Incongruent Equal Loadings Model

	Equal Loadings							
	White Men (N=11,108)		White Women (N=9,952)		Racially/ethnically Minoritized Men (N=2,364)		Racially/ethnically Minoritized Women (N=2,674)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	a	0.86***	a	0.85***	a	0.84***	a	0.85***
Content-Thought-Provoking	1.06***	0.83***	1.06***	0.80***	1.06***	0.80***	1.06***	0.81***
Material-Useful	1.15***	0.90***	1.15***	0.90***	1.15***	0.89***	1.15***	0.90***
Overall-Learning	0.42***	0.30***	0.42***	0.28***	0.42***	0.28***	0.42***	0.28***
Instructor								
Positive-Learning-Environment	1.34***	0.85***	1.34***	0.87***	1.34***	0.84***	1.34***	0.86***
Instructor-Organized	1.38***	0.82***	1.38***	0.84***	1.38***	0.83***	1.38***	0.83***
Instructor-Feedback	1.60***	0.87***	1.60***	0.89***	1.60***	0.84***	1.60***	0.89***
Overall-Learning	a	0.50***	a	0.51***	a	0.49***	a	0.53***
Mean Overall	a	a	a	a	a	a	a	a
Mean Instructor	a	a	a	a	a	a	a	a
R^2	0.981		0.981		0.977		0.982	
$\chi 2$	*df*=66, 672.80***							
CFI	0.995							
RMSEA	0.038							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p \leq 0.001$

A likelihood-ratio test was then used to compare the two models. The likelihood ratio test indicated that the invariant loading model performs statistically significantly worse than the same

form model ($\chi 2(18)=72.82$, $p \leq 0.001$). Thus, we should not constrain the loadings to be equal. This means that there are statistically significant differences between perceived gender role incongruent racially/ethnically minoritized women, racially/ethnically minoritized men, White women, and White men in the meaning of the latent variables Instructor and Overall when measured with the observed variables used in these models. Because the overall model fit of the metric invariance model is worse than the configural invariance model this means that at least one loading is not equivalent across the groups. Metric invariance is therefore *not* supported.

A postestimation test indicates that one of the observed variables in the model differs significantly between perceived gender role incongruent White men, White women, racially/ethnically minoritized men, and racially/ethnically minoritized women in the level of importance it carries in the measurement of the latent concepts. The results are presented in Table 22. There are significant chi-squared values for Content-Related-Assignments (25.656, $p<0.0001$), Content-Thought-Provoking (21.459, $p=0.0001$), and Overall-Learning (18.011, $p=0.0004$) on the latent concept Overall. This means that only these three variables which measure the latent concept Overall differ significantly on their levels of importance for gender role incongruent instructors based on their gender and race/ethnicity.

Table 22: Gender Role Incongruent Test of Group Invariance of Parameters

Latent Variables	Observed Variables	Score Test		
		χ2	*df*	*p-value*
Overall	Content-Related-Assignments	25.656	1	<0.0001
	Content-Thought-Provoking	21.459	1	0.0001
	Material-Useful	1.532	1	0.6750
	Overall-Learning	18.011	1	0.0004
Instructor	Positive-Learning-Environment	4.189	1	0.2417
	Instructor-Organized	2.099	1	0.5521
	Instructor-Feedback	1.728	1	0.6308
	Overall-Learning	2.737	1	0.4340

Since three variables were shown to be different across groups, a partial invariant loadings model can be run to compare the mean scores between the four perceived gender role incongruent instructor groups. In the partial invariant loadings model, the loadings for all of the variables which were determined to not be measured differently were constrained to be equal while the loadings for the three variables that were determined to be different were allowed to vary. The loadings for the partially invariant model were all positive but not statistically significant. The model fit statistics all indicated that the model fit well ($\chi 2(55)=260.10$, $p \leq 0.001$; CFI=0.996, TLI=0.993, SRMR=0.016, RMSEA=0.040). The full results of the partial invariant loadings model can be found in Table 23.

According to the partial invariant loadings model, Content-Related-Assignments on the latent concept Overall carries the most weight for perceived gender role incongruent White men (0.999) followed by White women (0.995), racially/ethnically minoritized men (0.992), and the least weight for racially/ethnically minoritized women (0.933). The variable Content-Thought-Provoking on the latent concept Overall carries the most weight for gender role incongruent racially/ethnically minoritized women (1.093) followed by White men (1.055), racially/ethnically minoritized men (1.044), and the least weight for White women (1.023). Finally, the variable Overall-Learning carries the most weight for perceived gender role incongruent racially/ethnically minoritized women (0.461) followed by White women (0.446), racially/ethnically minoritized men (0.405), and the least weight for White men (0.396).

Table 23: Gender Role Incongruent Partial Invariant Loadings Model

	Partial Invariant Loadings							
	White Men (*N*=11,108)		White Women (*N*=9,952)		Racially/ethnically Minoritized Men (*N*=2,364)		Racially/ethnically Minoritized Women (*N*=2,674)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	1.00	0.86***	1.00	0.85***	0.992	0.84***	0.93	0.83***
Content-Thought-Provoking	1.06	0.83***	1.02	0.79***	0.41	0.80***	1.09	0.83***
Material-Useful	1.14	0.90***	1.14	0.90***	1.14	0.89***	1.14	0.90***
Overall-Learning	0.40	0.28***	0.45	0.30***	0.96	0.27***	0.46	0.30***
Instructor								
Positive-Learning-Environment	0.96	0.85***	0.96	0.87***	0.96	0.84***	0.96	0.86***
Instructor-Organized	0.98	0.82***	0.98	0.84***	0.98	0.83***	0.98	0.83***
Instructor-Feedback	1.14	0.87***	1.14	0.89***	1.14	0.84***	1.14	0.89***
Overall-Learning	0.71	0.50***	0.71	0.50***	0.71	0.49***	0.71	0.52***
Mean Overall	a	a	a	a	a	a	a	a
Mean Instructor	a	a	a	a	a	a	a	a
R^2	0.981		0.981		0.977		0.982	
$\chi 2$	*df*=55, 620.10***							
CFI	0.996							
RMSEA	0.040							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Given the good fit statistics, the partially invariant model can be used to test for differences in the means between the groups, those being White men, White women, racially/ethnically minoritized men, and racially/ethnically minoritized women. The model was the same as the partial invariant loadings model but with the means allowed to vary rather than constrained. The results of the partially invariant means comparison model can be seen in table 10. The model fit well ($\chi 2(70)=1038.02$, $p\leq0.001$; CFI=0.992, TLI=0.991, SRMR=0.016, RMSEA=0.046) and the loadings were all positive, though not statistically significant. The full

results are presented in Table 24. The results indicate that there are *not* statistically significant differences between the scores of perceived gender role incongruent White men and the other three instructor groups—White women (Overall -0.14, p=0.970; Instructor -0.09, p=0.992), racially/ethnically minoritized men (Overall 0.07, p=0.970; Instructor 0.04, p=0.992), and racially/ethnically minoritized women (Overall -0.17, p=0.970; Instructor -0.22, p=0.992)—on the latent constructs Overall and Instructor[12]. These results indicate that when the model is properly constrained for measurement differences, there are *not* score differences on SEI forms based on instructor gender, race/ethnicity for perceived gender role incongruent instructors.

[12] The means comparison models were run with each of the four instructor groups as the reference group. The results of the other three rotations can be found in Appendices 7-9. All four of the models indicated that there were not statistically significant differences between any of the instructor groups on the means of either of the latent constructs. Since all of the models indicate the same results, only one is discussed in text.

Table 24: Role Incongruent Partial Invariant Loadings Model Means Comparison

	White Men as Reference Group							
	White Men (*N*=11,108)		White Women (*N*=9,952)		Racially/ethnically Minoritized Men (*N*=2,364)		Racially/ethnically Minoritized Women (*N*=2,674)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.79	0.86***	0.79	0.85***	0.79	0.84***	0.74	0.82***
Content-Thought-Provoking	0.84	0.83***	0.82	0.79***	0.83	0.80***	0.88	0.83***
Material-Useful	0.91	0.90***	0.91	0.90***	0.91	0.89***	0.91	0.90***
Overall-Learning	0.31	0.28***	0.37	0.31***	0.32	0.27***	0.37	0.30***
Instructor								
Positive-Learning-Environment	1.01	0.85***	1.01	0.87***	1.01	0.84***	1.01	0.86***
Instructor-Organized	1.04	0.82***	1.04	0.84***	1.04	0.82***	1.04	0.83***
Instructor-Feedback	1.20	0.87***	1.20	0.89***	1.20	0.84***	1.20	0.89***
Overall-Learning	0.75	0.50***	0.75	0.50***	0.75	0.49***	0.75	0.52***
Mean Overall	a	a	-0.14	-0.14***	0.07	0.08***	-0.17	-0.18***
Mean Instructor	a	a	-0.09	-0.12***	0.04	0.06***	-0.22	-0.27***
R^2	0.981		0.980		0.977		0.982	
χ2	*df*=70, 1038.02***							
CFI	0.992							
RMSEA	0.046							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Gender Role Neutral Faculty

Grouped Structural Equation Models

A grouped SEM was conducted utilizing the CFA model from Study 1 (Chapter 2) and grouped based on instructor gender and race. To test for configural invariance, a same form equivalence model was run in which the means of the latent concepts were set to equal zero but there were no constraints placed on the groups. The loadings were all substantial and statistically significant and the model fit statistics all indicated that the model fit well ($\chi 2(48)=376.58$,

$p \leq 0.001$; CFI=0.995, TLI=0.991, SRMR=0.015, RMSEA=0.047). The results of the same form equivalence model for perceived gender role neutral can be found in Table 25.

Table 25: Gender Role Neutral Same Form Equivalence Model

	Same Form Equivalence							
	White Men (N=4,889)		White Women (N=6,018)		Racially/ethnically Minoritized Men (N=944)		Racially/ethnically Minoritized Women (N=791)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	a	0.85***	a	0.85***	a	0.79***	a	0.83***
Content-Thought-Provoking	1.04***	0.82***	1.08***	0.81***	0.93***	0.73***	1.07***	0.82***
Material-Useful	1.13***	0.90***	1.18***	0.91***	1.12***	0.87***	1.16***	0.88***
Overall-Learning	0.40***	0.26***	0.51***	0.33***	0.70***	0.45***	0.47***	0.29***
Instructor								
Positive-Learning-Environment	1.14***	0.84***	1.37***	0.85***	1.74***	0.82***	1.08***	0.83***
Instructor-Organized	1.30***	0.82***	1.36***	0.82***	1.59***	0.77***	1.23***	0.82***
Instructor-Feedback	1.40***	0.88***	1.64***	0.90***	2.10***	0.87***	1.37***	0.89***
Overall-Learning	a	0.56***	a	0.53***	a	0.40***	a	0.60***
Mean Overall	a	a	a	a	a	a	a	a
Mean Instructor	a	a	a	a	a	a	a	a
R^2	0.977		0.982		0.967		0.974	
$\chi 2$	*df*=48, 376.58***							
CFI	0.995							
RMSEA	0.047							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p \leq 0.001$

To test for metric invariance, an equal loadings model was run in which the means of the latent concepts were still set to equal zero and the measurement coefficients were constrained to be equal across the groups. The equal loadings model for perceived gender role neutral can be found in Table 26. The loadings were once again all substantial and statistically significant and

the model fit statistics all indicated that the model fit well ($\chi 2(66)=497.81$, $p \leq 0.001$; CFI=0.993, TLI=0.992, SRMR=0.036, RMSEA=0.046).

Table 26: Gender Role Neutral Equal Loadings Model

	Equal Loadings							
	White Men (*N*=4,889)		White Women (*N*=6,018)		Racially/ethnically Minoritized Men (*N*=944)		Racially/ethnically Minoritized Women (*N*=791)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	a	0.85***	a	0.85***	a	0.78***	a	0.84***
Content-Thought-Provoking	1.06***	0.82***	1.06***	0.81***	1.06***	0.76***	1.06***	0.82***
Material-Useful	1.16***	0.90***	1.16***	0.91***	1.16***	0.87***	1.16***	0.88***
Overall-Learning	0.48***	0.32***	0.48***	0.31***	0.48***	0.31***	0.48***	0.31***
Instructor								
Positive-Learning-Environment	1.29***	0.85***	1.29***	0.84***	1.29***	0.80***	1.29***	0.85***
Instructor-Organized	1.35***	0.80***	1.35***	0.82***	1.35***	0.80***	1.35***	0.81***
Instructor-Feedback	1.57***	0.89***	1.57***	0.89***	1.57***	0.85***	1.57***	0.89***
Overall-Learning	a	0.51***	a	0.55***	a	0.52***	a	0.56***
Mean Overall	a	a	a	a	a	a	a	a
Mean Instructor	a	a	a	a	a	a	a	a
R^2	0.977		0.982		0.966		0.975	
$\chi 2$	*df*=66, 497.81***							
CFI	0.993							
RMSEA	0.046							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p \leq 0.001$

A likelihood-ratio test was then used to compare the two models. The likelihood ratio test indicated that the invariant loading model performs statistically significantly worse than the same form model ($\chi 2(18)=121.23$, $p \leq 0.001$). Thus, we should not constrain the loadings to be equal. This means that there are statistically significant differences between perceived gender role

neutral racially/ethnically minoritized women, racially/ethnically minoritized men, White women, and White men in the meaning of the latent variables Instructor and Overall when measured with the observed variables used in these models. Because the overall model fit of the metric invariance model is worse than the configural invariance model this means that at least one loading is not equivalent across the groups. Metric invariance is therefore *not* supported.

A postestimation test indicates that three of the observed variables in the model differ significantly between perceived gender role neutral White men, White women, racially/ethnically minoritized men, and racially/ethnically minoritized women in the level of importance they carry in their measurement of the latent concepts. The output which can be seen in Table 27 shows significant chi-squared values for Content-Related-Assignments (9.629, p=0.0220), Content-Thought-Provoking (12.482, p=0.0059), and Overall-Learning (15.706, p=0.0013) on the latent concept Overall and Positive-Learning-Environment (26.710, $p \leq 0.001$), Instructor-Organized (64.018, $p \leq 0.001$), Instructor-Feedback (10.406, p=0.0154), and Overall-Learning (14.677, p=0.0021) on the latent concept Instructor. This means that all but one variable, Material-Useful (3.407, p=0.3330), differs significantly on their levels of importance for perceived gender role neutral instructors based on their gender and race/ethnicity.

Table 27: Gender Role Neutral Test of Group Invariance of Parameters

Latent Variables	Observed Variables	Score Test		
		χ2	*df*	*p-value*
Overall	Content-Related-Assignments	9.629	1	0.0220
	Content-Thought-Provoking	12.482	1	0.0059
	Material-Useful	3.407	1	0.3330
	Overall-Learning	15.706	1	0.0013
Instructor	Positive-Learning-Environment	26.710	1	<0.001
	Instructor-Organized	64.018	1	<0.001
	Instructor-Feedback	10.406	1	0.0154
	Overall-Learning	14.677	1	0.0021

A partial invariant loadings model can be run to compare the mean scores between White men, White women, racially/ethnically minoritized men, and racially/ethnically minoritized women. In the partial invariant loadings model, the loadings for the one variable which was determined to not be measured differently was constrained to be equal while the loadings for the seven variables that were determined to be different were allowed to vary. Since all of the variables measuring the latent concept Instructor were determined to be measured differently, one of the variables needs to be constrained in order to run the partial invariant model. To test for differences in the constrained variable, the constrained variable will be rotated in order to test for differences in every variable. The loadings for the partially invariant model were all positive but not statistically significant and can be found in Tables 28 and 29. The model fit statistics all indicated that the model fit well ($\chi 2(47)=376.58$, $p \leq 0.001$; CFI=0.995, TLI=0.991, SRMR=0.015, RMSEA=0.047).

According to the partial invariant loadings model, Content-Related-Assignments on the latent concept Overall carries the least weight for perceived gender role neutral White women (0.90) followed by racially/ethnically minoritized women (0.92), White men (0.93), and the most weight for racially/ethnically minoritized men (0.94). Content-Thought-Provoking on the latent concept Overall carries the least weight for perceived gender role neutral racially/ethnically minoritized men (0.88) followed by White men (0.97)[13], White women (0.97), and the most weight for racially/ethnically minoritized women (0.98). Overall-Learning on the latent concept Overall carries the least weight for gender role neutral White men (0.37) followed by racially/ethnically minoritized women (0.43), White women (0.45), and the most weight for racially/ethnically minoritized men (0.66). Positive-Learning-Environment on the latent concept

[13] The full reported value was considered when selecting which was higher. In this case, the coefficient for White men was 0.966 and for racially/ethnically minoritized women it was 0.972.

Instructor carries the least weight for perceived gender role neutral racially/ethnically minoritized women (1.08) followed by White men (1.14), White women (1.37), and the most weight for racially/ethnically minoritized men (1.74). Overall-Learning on the latent concept Instructor carries the least weight for perceived gender role neutral racially/ethnically minoritized men (0.57) followed by White women (0.73), White men (0.88), and the most weight for racially/ethnically minoritized women (0.92). All of the patterns listed were consistent regardless of which variable was constrained on the Instructor latent concept.

The weights for the variables Instructor-Organized and Instructor-Feedback were more susceptible to variation based on which other variable was constrained in the model. When Overall-Learning on the latent concept Instructor was constrained, Instructor-Organized carries the least weight for perceived gender role neutral racially/ethnically minoritized women (1.13) followed by White men (1.30), White women (1.59), and the most weight for racially/ethnically minoritized men (1.59). When Positive-Learning-Environment was constrained, Instructor-Organized carries the least weight for perceived gender role neutral racially/ethnically minoritized men (0.91) followed by White women (1.00), racially/ethnically minoritized women (1.13), and the most weight for White men (1.14).

The variable Instructor-Feedback, when Overall-Learning on the latent concept Instructor is constrained, carries the least weight for perceived gender role neutral racially/ethnically minoritized women (1.37) followed by White men (1.40), White women (1.64), and the most weight for racially/ethnically minoritized men (2.10). When the variable Positive-Learning-Environment is constrained, Instructor-Feedback carries the least weight for perceived gender role neutral White women (1.20) followed by racially/ethnically minoritized men (1.21), White men (1.22), and the most weight for racially/ethnically minoritized women (1.27).

Table 28: Gender Role Neutral Partial Invariant Model

	Overall-Learning Constrained							
	White Men (*N*=4,889)		White Women (*N*=6,018)		Racially/ethnically Minoritized Men (*N*=944)		Racially/ethnically Minoritized Women (*N*=791)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.93	0.85***	0.90	0.85***	0.94	0.79***	0.92	0.83***
Content-Thought-Provoking	0.97	0.82***	0.97	0.81***	0.88	0.73***	0.98	0.82***
Material-Useful	1.06	0.90***	1.06	0.91***	1.06	0.87***	1.06	0.88***
Overall-Learning	0.37	0.26***	0.45	0.33***	0.66	0.45***	0.43	0.29***
Instructor								
Positive-Learning-Environment	1.14***	0.84***	1.37***	0.85***	1.74***	0.82***	1.08***	0.83***
Instructor-Organized	1.30***	0.82***	1.36***	0.82***	1.59***	0.77***	1.23***	0.82***
Instructor-Feedback	1.40***	0.88***	1.64***	0.90***	2.10***	0.87***	1.37***	0.89***
Overall-Learning	a	0.56***	a	0.53***	a	0.40***	a	0.60***
Mean Overall	a	a	a	a	a	a	a	a
Mean Instructor	a	a	a	a	a	a	a	a
R^2	0.977		0.982		0.967		0.974	
$\chi 2$	*df*=47, 376.58***							
CFI	0.995							
RMSEA	0.047							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Table 29: Gender Role Neutral Partial Invariant Model Rotating Constraints

	Positive-Learning-Environment Constrained							
	White Men (*N*=4,889)		White Women (*N*=6,018)		Racially/ethnically Minoritized Men (*N*=944)		Racially/ethnically Minoritized Women (*N*=791)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.87	0.85***	0.84	0.85***	0.88	0.79***	0.85	0.83***
Content-Thought-Provoking	0.90	0.82***	0.91	0.81***	0.82	0.73***	0.91	0.82***
Material-Useful	0.99	0.90***	0.99	0.91***	0.99	0.87***	0.99	0.88***
Overall-Learning	0.35	0.26***	0.42	0.33***	0.62	0.45***	0.40	0.29***
Instructor								
Positive-Learning-Environment	a	0.84***	a	0.85***	a	0.82***	a	0.83***
Instructor-Organized	1.14***	0.82***	1.00***	0.82***	0.91***	0.77***	1.13***	0.82***
Instructor-Feedback	1.22***	0.88***	1.20***	0.90***	1.21***	0.87***	1.27***	0.89***
Overall-Learning	0.88***	0.56***	0.73***	0.53***	0.57***	0.40***	0.92***	0.60***
Mean Overall	a	a	a	a	a	a	a	a
Mean Instructor	a	a	a	a	a	a	a	a
R^2	0.977		0.982		0.967		0.974	
$\chi 2$	*df*=47, 376.58***							
CFI	0.995							
RMSEA	0.047							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*p<0.05; **p<0.01; ***p≤0.001

Given the good fit statistics, the partially invariant model can be used to test for differences in the means between the groups, those being racially/ethnically minoritized women, racially/ethnically minoritized men, White men, and White women. The partially invariant means comparison model results for perceived gender role neutral instructors are presented in Tables 30 and 31. The model fit well ($\chi 2(62)=475.27$, $p\leq 0.001$; CFI=0.994, TLI=0.991, SRMR=0.020, RMSEA=0.046) and the loadings were all positive, though not statistically significant. The results indicate that there are not statistically significant differences between the

means of perceived gender role neutral White men and the other three instructor groups —White women (-0.10, p=0.954), racially/ethnically minoritized men (-0.30, p=0.954), and racially/ethnically minoritized women (-0.28, p=0.954)—on the latent construct Overall. These results indicate that when the model is properly constrained for measurement differences, there are not score differences on the observed variables for the latent construct Overall based on instructor gender, race/ethnicity, and perceived gender role neutrality.

There are significant differences between the means of the different instructor groups for the latent concept Instructor. The differences in the means for the Instructor latent concept vary depending on which instructor group is the reference group. The results in Tables 18 and 19 use White men as the reference group. The remaining reference group rotations can be found in Appendices 10-15. There are some general patterns to the results such that the means for perceived gender role neutral White men are statistically significantly higher than of any other instructor groups. Perceived gender role neutral White women have statistically significantly higher means than racially/ethnically minoritized men and racially/ethnically minoritized women but lower means than White men. Perceived gender role neutral racially/ethnically minoritized men have statistically significantly lower means than all other groups. Perceived gender role neutral racially/ethnically minoritized women also have statistically significantly lower means than White men and White women but there is not a statistically significant relationship when racially/ethnically minoritized men are the reference group. In summary, the means of perceived gender role neutral White men are statistically significantly higher than all other groups while the means of racially/ethnically minoritized men are statistically significantly lower than all other groups. The mean scores of perceived gender role neutral White women are statistically significantly higher than all groups except White men and the scores of White women are

statistically significantly lower than White men and White women among perceived gender role neutral faculty.

Table 30: Gender Role Neutral Partial Invariant Model Means Comparison

	Overall-Learning on Instructor Constrained							
	White Men (*N*=4,889)		White Women (*N*=6,018)		Racially/ethnically Minoritized Men (*N*=944)		Racially/ethnically Minoritized Women (*N*=791)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.99	0.85***	0.95	0.85***	1.02	0.80***	0.96	0.83***
Content-Thought-Provoking	1.02	0.82***	1.03	0.82***	0.93	0.72***	1.03	0.82***
Material-Useful	1.12	0.90***	1.12	0.91***	1.12	0.86***	1.12	0.88***
Overall-Learning	0.39	0.26***	0.48	0.33***	0.66	0.41***	0.43	0.27***
Instructor								
Positive-Learning-Environment	1.14***	0.84***	1.38***	0.85***	1.50***	0.82***	1.02***	0.82***
Instructor-Organized	1.30***	0.82***	1.36***	0.81***	1.27***	0.74***	1.19***	0.82***
Instructor-Feedback	1.40***	0.88***	1.65***	0.90***	1.81***	0.87***	1.31***	0.89***
Overall-Learning	a	0.56***	a	0.53***	a	0.45***	a	0.62***
Mean Overall	a	a	-0.10	-0.13***	-0.30	-0.38***	-0.28	-0.33***
Mean Instructor	a	a	-0.11***	-0.18***	-0.30***	-0.52***	-0.38***	-0.45***
R^2	0.977		0.982		0.967		0.974	
$\chi 2$	*df*=62, 475.27***							
CFI	0.994							
RMSEA	0.046							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Table 31: Gender Role Neutral Partial Invariance Model Means Comparison Rotating Constraints

	Positive-Learning-Environment Constrained							
	White Men (*N*=4,889)		White Women (*N*=6,018)		Racially/ethnically Minoritized Men (*N*=944)		Racially/ethnically Minoritized Women (*N*=791)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	1.01	0.85***	0.97	0.85***	1.05	0.80***	0.98	0.83***
Content-Thought-Provoking	1.05	0.82***	1.05	0.82***	0.95	0.72***	1.05	0.82***
Material-Useful	1.14	0.90***	1.14	0.91***	1.14	0.86***	1.14	0.88***
Overall-Learning	0.40	0.26***	0.49	0.33***	0.67	0.41***	0.44	0.27***
Instructor								
Positive-Learning-Environment	a	0.84***	a	0.85***	a	0.82***	a	0.82***
Instructor-Organized	1.14***	0.82***	0.99***	0.81***	0.85***	0.74***	1.16***	0.82***
Instructor-Feedback	1.23***	0.88***	1.20***	0.90***	1.21***	0.87***	1.28***	0.89***
Overall-Learning	0.88***	0.56***	0.73***	0.53***	0.67***	0.45***	0.98***	0.62***
Mean Overall	a	a	-0.10	-0.13***	-0.29	-0.38***	-0.28	-0.33***
Mean Instructor	a	a	-0.16***	-0.18***	-0.45***	-0.52***	-0.39***	-0.78***
R^2	0.977		0.982		0.967		0.974	
$\chi 2$	*df*=62, 475.27***							
CFI	0.994							
RMSEA	0.046							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*p<0.05; **p<0.01; ***p≤0.001

DISCUSSION

The results of the grouped structural equation modeling varied greatly by the gender, race/ethnicity, and level of perceived role (in)congruity of the instructor. On the whole, these results, much like the results of Study 1 (Chapter 2), illustrate the importance of determining the appropriateness of a measurement model prior to testing for differences in the mean student

evaluation scores for faculty of different genders, races/ethnicities, and perceived level of role congruity. Furthermore, the results of this study indicate that even when measurement invariance is accounted for, some of the between group differences in SEI scores persist depending on the perceived role congruity of the group being examined.

For perceived role congruent faculty, even once the model was properly constrained according to the results of measurement invariance testing, some of the statistically significant differences in the means of the instructor groups persisted. Among perceived gender role congruent faculty, the means for the latent concept Overall for White women was higher than all other faculty groups. The means for the latent concept Overall for racially/ethnically minoritized women were lower than all other faculty groups. This indicates that gender and race/ethnicity interact with respect to evaluations of perceived role congruent faculty such that White women are evaluated more highly than all others while racially/ethnically minoritized women are evaluated more negatively than all others. Thus, while White women are evaluated the most positively when they teach in woman-dominated courses, racially/ethnically minoritized women are evaluated the most positively. The discrepancy in evaluations of perceived role congruent women may be due to the fact that while the disciplines in which they teach are historically woman-dominated they are historically White woman-dominated. Racially/ethnically minoritized women may not be perceived as feminine enough to properly take on the role of woman instructor of a woman-dominated course because of their racial/ethnic difference from the traditional majority of instructors. While racially/ethnically minoritized women occasionally benefit from not being viewed as feminine as White women (Livingston et al. 2012), this is not the case when examining the student evaluations of perceived role congruent women-instructors.

Additionally on the latent concept Overall, White perceived gender role congruent men are evaluated more positively than racially/ethnically minoritized men and racially/ethnically minoritized women. Racially/ethnically minoritized men are evaluated more negatively than all groups except racially/ethnically minoritized women. These two results further indicate that gender and race/ethnicity affect student evaluations of instruction such that White persons are at an advantage relative to racially/ethnically minoritized persons in perceived gender role congruent positions. Racially/ethnically minoritized men may be penalized like their racially/ethnically minoritized women peers for not fitting with the traditionally White men associated with teaching in traditionally man-dominated disciplines. Thus, racially/ethnically minoritized faculty even when they are perceived to be role congruent are seen as lacking fit with their occupation and thus receive lower evaluations than their role congruent White peers.

There were not statistically significant differences on the means of the latent concept Instructor for gender role congruent faculty. Thus, student evaluations of the overall quality of their course for perceived role congruent faculty is more affected by the gender and race/ethnicity of the instructors than student evaluations of the instructors themselves. Students' evaluations of their instructors may vary more depending on the gender and race/ethnicity of their instructor with respect to course-related and not instructor-related items for fear of appearing biased, racist, and/or sexist. Students may believe that evaluating racially/ethnically minoritized faculty and especially racially/ethnically minoritized women harshly on questions related to them as an instructor and therefore person may be perceived negatively while evaluating aspects of the course negatively does not carry the same potential connotations. Thus, students' unconscious biases against racially/ethnically minoritized men and especially racially/ethnically minoritized women instructors come through more so in questions about

course content than in questions about the instructor. To overcome these biases, a neutral third party could evaluate the course materials of instructors of different races/ethnicities to determine if they are related to assignments, thought-provoking, and/or useful however this will not help in overcoming students' unconscious biases again the course materials of their racially/ethnically minoritized instructors.

For perceived gender role incongruent faculty, once the models were constrained based on the determined measurement invariance there were not statistically significant differences in the means of either latent construct, Overall and Instructor. This finding indicates that all perceived role incongruent instructor groups regardless of their gender or race/ethnicity are evaluated similarly on questions asking about the instructor or the course materials. Accounting for measurement invariance eliminated all gender and race/ethnicity based evaluation differences among perceived gender role incongruent faculty but prior to adding these constraints, there were differences on three of the observed variables for the latent concept Overall: Content-Related-Assignments, Content-Thought-Provoking, and Overall-Learning.

Though these between group differences were accounted for in the partial invariant loadings model, for perceived role incongruent faculty there are more between group differences in the measurement of students' overall learning than there are in the measurement of students' evaluation of their instructors. Therefore, if student evaluation scores of perceived role incongruent faculty are compared to one another without measurement invariance accounted for, there will be larger differences between the scores of different instructor groups on the latent construct Overall than on the latent construct Instructor. This finding indicates that students are more likely to evaluate instructors they perceive to be role incongruent of different genders and/or race/ethnicities differently on concepts related more to the course than to the instructor

themselves. Much like with perceived gender role congruent racially/ethnically minoritized instructors, students may be hesitant to critique role incongruent instructors on the basis of them as an instructor for fear of being seen as being biased against people who violate gender norms and thus provide their more divisive feedback with respect to the course. Again, course materials may be able to be independently evaluated however the problem will persist in which students perceive some instructors to have more useful, thought-provoking, and/or related assignments than others. Students would need to be educated about unconscious biases and willing to take a critical look at their schematic processing in order to truly change their perceptions and evaluations of instructors of different genders and/or race/ethnicities.

For perceived gender role neutral faculty, much like perceived gender role congruent faculty even once the model was properly constrained according to the results of measurement invariance testing, some of the statistically significant differences in the means of the instructor groups persisted. There were not persistent statistically significant differences between the means of the instructor groups on the latent concept Overall. Among perceived gender role neutral faculty, there were statistically significant differences that persisted between the means of the instructor groups on the latent concept Instructor. The means for the latent concept Instructor for White men were statistically significantly higher than all other instructor groups. The means for racially/ethnically minoritized men were statistically significantly lower than all other instructor groups. These juxtaposed results for men indicate that the race/ethnicity of the instructor has a large effect on students' evaluations of instruction. This finding supports previous research which has found that student evaluations tend to be biased against racially/ethnically minoritized instructors (Reid 2010; Smith and Hawkins 2011).

Furthermore, the means for White women were statistically significantly lower than White men but higher than those of both racially/ethnically minoritized men and women. Finally, the means for racially/ethnically minoritized women were statistically significantly lower than that of White men and White women. The pattern of results among women further highlights that the race/ethnicity of the instructor has an impact on students' evaluations of instruction that is in line with previous findings such that White instructors are likely to receive higher evaluations than racially/ethnically minoritized instructors (Aruguete et al. 2017; Smith and Hawkins 2011). White instructors in gender role neutral disciplines and especially White men are at an advantage relative to their racially/ethnically minoritized and women counterparts, a result which is also supported by previous research (Chávez and Mitchell 2020). While these results support the findings of previous work such that White instructors and White men in particular are more likely to receive higher evaluation scores, they add another dimension to prior research as the results of this study indicate that this is the case in gender role neutral disciplines. For perceived role congruent faculty, White women received the highest scores, a finding which is in contrast to previous literature (Boring et al. 2016; MacNell et al. 2015a). Thus these results indicate that it is important to consider not only the gender but also the perceived gender role congruity, and the race/ethnicity of instructors when comparing student evaluation scores.

These results add another dimension to the results of the analyses in Study 1 (Chapter 2). In Study 1 (Chapter 2), once the models accounted for measurement invariance there were no longer statistically significant differences between men and women instructors in any of the three perceived role congruity groups. The results of this study indicate that when both instructor gender and race/ethnicity are accounted for, some of the statistically significant differences among perceived gender role congruent and perceived gender role neutral faculty persist even

after measurement invariance is taken into consideration. For perceived gender role congruent faculty, statistically significant differences on the latent concept Overall persisted even when measurement invariance was accounted for. For perceived gender role neutral faculty, statistically significant differences on the latent concept Instructor persisted even when measurement invariance is accounted for.

When instructors teach in disciplines in which they are a part of the gender dominant group, students' perceptions of their Overall learning in the course are affected by the gender and race/ethnicity of their instructors such that White women receive higher scores than the other instructor groups. This may be because White women are seen as experts in women-dominated disciplines in such a way that students perceive their Overall learning to be higher than men teaching in man-dominated disciplines. But White women may also be at an advantage in women-dominated disciplines over White men in man-dominated disciplines because students tend to rate man-dominated classes such as science and math classes lower than women-dominated classes such as English (Basow and Montgomery 2005; Wachtel 1998). However, the advantage White women receive when teaching in women-dominated disciplines does not extend to racially/ethnically minoritized women whose scores were statistically significantly lower than all other groups. These results indicate that while White women are at an advantage in women-dominated disciplines, racially/ethnically minoritized women are still penalized due to their race. Additionally, while it appears that White women may be at an advantage in perceived role congruent disciplines, they may be putting in more effort and going above and beyond their job description in order to receive marginally higher evaluation scores than their White man counterparts (El-Alayli et al. 2018). An examination of the time spent on student/teaching related tasks among role congruent faculty would be necessary to determine if White role congruent

women are truly at an advantage or if they simply put in more time and effort than their White perceived role congruent men counterparts.

When instructors teach in disciplines in which there is about an equal distribution of men and women (perceived gender role neutral), students' perceptions of their Instructor is affected by the gender but more so the race/ethnicity of their instructors such that White men receive higher scores than all other groups and White women receive higher scores than racially/ethnically minoritized men and women instructors. Thus, in perceived gender role neutral disciplines students tend to rate White instructors more positively on items that ask about if the instructor created a positive learning environment, if the instructor was organized, if the instructor provided good feedback, and if their overall learning in the course was good. When there is not a clear gender-dominance in a course, it seems that students rely on their unconscious racial/ethnic biases such that White persons are rated more positively on measures specifically related to the instructor of the course than their racially/ethnically minoritized counterparts. To my knowledge, prior research has not parsed out student evaluation scores based on the types of questions asked thus this finding constitutes a novel contribution to the area of study. Researchers and institutions need to consider how specific observed variables combine to measure different latent constructs when studying student evaluations of instruction.

Taken together, these results indicate that instructor gender, race/ethnicity, and perceived level of gender role congruity do affect student evaluations of instruction. For gender role congruent and role neutral faculty, statistically significant differences between instructor groups persisted even after the models were properly constrained for measurement invariance. The persistent statistical differences are such that White faculty receive higher scores than their racially/ethnically minoritized counterparts. Among perceived role congruent faculty on the

latent concept Overall, White women receive an even higher score than White men while among role neutral faculty White men receive an even higher score than White women. For racially/ethnically minoritized role congruent faculty, racially/ethnically minoritized women receive lower scores than racially/ethnically minoritized men while for role neutral faculty, racially/ethnically minoritized men receive lower scores than racially/ethnically minoritized women. These results indicate that while White instructors are at an advantage relative to their racially/ethnically minoritized counterparts, this advantage varies by the gender and perceived level of gender role congruity of the instructor.

Thus, if institutions of higher education seek to equitably evaluate their instructors it is important for them to consider not only the gender but the race/ethnicity, and the perceived level of role congruity of the instructor when constructing and analyzing student evaluations of instruction. Furthermore, it is important for institutions to complete rigorous testing such as Confirmatory Factor Analyses and Grouped Structural Equation Models to determine if the observed variables are measuring the latent constructs they seek to measure and to determine if and account for as much residual measurement invariance between different instructor groups as possible. If institutions are thoughtful in the design and analysis of their student evaluations of instruction, they can ensure they are measuring what they desire to measure and do so in as equitable of a manner as possible.

Limitations

This study is not without its limitations. Sample sizes of race/ethnicity groups other than White were not large enough for statistical power thus requiring all racially/ethnically minoritized identities to be combined into one category. While this combining allowed for statistically large enough sample sizes, it is oversimplifying the unique experiences of each

separate racial/ethnic group by assuming their experiences are all the same simply because they are not White. As shown by the results of this study, the unique combination of instructor identities can affect the ways in which students evaluate their instructors. Therefore, the unique racially/ethnically minoritized identity of a person may affect students' perceptions in a way that is unique from other racially/ethnically minoritized persons. Future studies should seek to have greater sample sizes within each racial/ethnic category so that they do not need to be merged into one group for statistical power.

Additionally, even with the merging of all racially/ethnically minoritized persons into one category, there were still very small racially/ethnically minoritized response sizes for faculty in role-neutral disciplines (944 men, 791 women). While there were enough cases for statistical power (Kline 2015), a more equitable sample size to that of the White role-neutral persons (4,889 men, 6,018 women) may change the results of the study. Future studies should seek to have more representation of instructors in role-neutral fields if possible.

While these are limitations of this study, they were caused by the population of instructors at the institution being studied. This speaks to a larger representation problem in academia which is still heavily dominated by White persons in general and White men in particular. Institutions of higher education should examine the current gender, racial/ethnic, and disciplinary diversity of their faculty and take any disparities into consideration when hiring new faculty. Through instructor diversity enhancement programs, issues of discrepant evaluations of faculty may dissipate as more diverse faculty become more normal throughout the academe. Additionally, increasing faculty diversity may help to encourage other marginalized persons to pursue avenues they have been historically excluded from thus helping to enhance diversity in many historically White man dominated domains such as STEM fields and business.

Additionally, as described in Study 1 (Chapter 2) the data tends to skew relatively high. The mean scores of the seven variables ranged from 3.897 to 4.419 on a five-point scale. These averages indicate that while there is some variation in student evaluation scores, students generally rate their instructors positively on quantitative SEI measures. Institutions of higher education may want to reconsider the questions they are asking on their student evaluation forms if the goal is to distinguish between the most effective and the least effective instructors regardless of their gender, race/ethnicity, and the discipline in which they teach.

Future Studies

While the inclusion of three levels of instructor identity is an important step towards more intersectional studies of student evaluations of instruction, there are many more identities of instructors, courses, and students which may affect the ways in which students evaluate their instructors. For example, the temperament of the instructor was not included in the analyses but may affect students' perceptions of them and the course(s) that they teach. Future studies may want to consider additional faculty characteristics such as temperament, sexuality, level of experience, etc. Furthermore, the gender, race/ethnicity, and major discipline of the students completing the evaluations were not considered in this study. According to congruity theories, a person's own identities and especially their own level of gender role (in)congruity may affect the ways in which they evaluate others (Diekman and Schneider 2010b; Eagly and Karau 2002). Future studies should consider student identities as well as more instructor identities in their examination of student evaluations of instruction.

Additionally, while this study provides a richer analysis of student evaluations of instruction, the causal mechanisms behind the persistent gender and racial/ethnic differences are still unclear. Two of the leading theories of congruity posit different arguments as to why

backlash to role incongruity occurs. Role Congruity Theory (RCT) posits that perceived role incongruity between the salient social role gender and other salient social roles such as occupation leads to backlash (Diekman and Eagly 2008; Eagly et al. 2000). Status Incongruity Hypothesis (SIH) posits that a visceral reaction to deviations from the gender hierarchy leads to backlash (Rudman and Glick 2001b). The specific causal mechanisms behind the differences in student evaluation scores could not be determined in this study but a qualitative study of the open-response questions on student evaluation forms may help to determine the causality behind score differences. Future studies should examine open-response student evaluation questions while considering the gender, race/ethnicity, and role congruity of the instructor being evaluated.

CONCLUSION

The results of this study in combination with Study 1 (Chapter 2) indicate the importance of testing and accounting for measurement invariance among diverse groups of targets. Measurement invariance is a crucial step to conduct prior to any other between group testing to be sure that all groups being analyzed are on the same playing field. Every study and evaluation should be sure to test for measurement invariance before conducting significance testing. Furthermore, these results in tandem with Study 1 (Chapter 2) indicate that while statistically significant differences between men and women instructors are accounted for in every role congruity group when measurement invariance is taken into consideration, this is not the case when instructor gender and race/ethnicity are also considered. When both gender and race/ethnicity are considered and measurement invariance is accounted for, statistically significant differences on the Overall latent construct for perceived gender role congruent instructors and statistically significant differences on the Instructor latent construct for perceived

gender role neutral instructors persists such that White instructors are advantaged relative to their racially/ethnically minoritized counterparts.

These results indicate the importance not only of accounting for measurement invariance but also for considering the multiple intersecting identities of instructors when studying student evaluations of instruction. This raises the question of what the causal mechanisms behind these gender, racial/ethnic, and perceived gender role congruity differences in student evaluations of instruction are. One possible way to test for causal mechanisms behind these differences is to examine the ways in which students talk about their instructors on the open-response student evaluation questions. In the next study, I examine the ways in which students talk about their instructors on open-response student evaluation questions and how these responses vary by instructor gender, race/ethnicity, and/or perceived gender role congruity to discern the causal mechanisms behind differences in student evaluation scores.

CHAPTER 4: QUALITATIVE ANALYSES OF STUDENT EVALUATIONS OF INSTRUCTION WITH ATTENTION TO FACULTY RACE/ETHNICITY, GENDER, AND GENDER ROLE (IN)CONGRUITY

INTRODUCTION

Subjective evaluations are evaluations based on opinions and perceptions rather than based on objective facts. Student evaluations of instruction (SEI) are one form of subjective evaluation which have been studied thoroughly. Research has consistently found that student evaluations tend to be biased against women and faculty of color (Arbuckle and Williams 2003; Kobrynowicz and Biernat 1997; Liden, Stilwell, and Ferris 1996; Smith et al. 2001; Smith et al. 2019). As established in the previous three chapters, there are many factors that can affect students' subjective evaluations of their instructors such as the gender, race/ethnicity, and/or discipline/field of the instructor (Anon 2019; Basow 1995; Bavishi et al. 2010). The previous two studies stand somewhat in contrast to previous research on student evaluations. The results of the previous two studies indicate that when measurement invariance is accounted for, most differences in evaluation scores based on instructor gender, race/ethnicity, and perceived gender role congruity are minimized or even eliminated.

In the previous two quantitative studies of student evaluations, it was found that the unique intersections of multiple instructor identities do affect student evaluations of their instructors. While these results are compelling, there is much anecdotal and scientific evidence that suggests that women faculty tend to receive different types of qualitative comments from students than men faculty which were not captured in the previous two studies of quantitative student evaluation measures (Falkoff 2018; McMurtrie 2019; Mitchell and Martin 2018). A purely quantitative analysis of student evaluations may not be able to ascertain the causal

mechanisms behind any differences in students' perceptions and evaluations of their instructors based on their gender, perceived gender role (in)congruity, and/or race/ethnicity.

Additionally, the gendered and/or racialized expectations of students regarding teaching style and additional task expectations of their instructors may not come through in the quantitative SEI questions alone. These additional burdens placed on women and racially/ethnically minoritized faculty may lead to higher burnout rates and more time spent on non-career-enhancing tasks such as more special favor requests and reciprocation of friendship behaviors to appease students by meeting their status expectations which may in turn lead to higher student evaluations (El-Alayli et al. 2018). A purely quantitative analysis of SEIs has the potential to miss both the effects of students' unconscious biases and their additional expectations for women and racially/ethnically minoritized faculty which may only come through in their written comments. The results of Studies 1 and 2 (Chapters 2 and 3) indicate that perceived role congruent women and perceived role congruent white women in particular, actually receive higher evaluations than their male counterparts when accounting for measurement invariance. While these results are interesting, women's scores may be inflated due to the extra tasks they are completing but there is no way to know if this is the case from the quantitative analyses alone. The quantitative questions examined in Studies 1 and 2 (Chapters 2 and 3) ask questions about both the course and the instructor, themes which carry over into the qualitative questions examined in this study. Thus, in order to contextualize the quantitative measures it is important to also consider students' open-response student evaluation questions as they provide students an opportunity to express their opinions about the instructor and/or course free from the restrictions of quantitative scales which allows students to also talk about the out-of-class work their instructors may be doing for them.

In this study, student evaluations of instruction are qualitatively analyzed with consideration of the gender, perceived gender role (in)congruity, and race/ethnicity of the course instructor. Through these analyses, I seek to answer the questions: are students' subjective evaluations of their instructors affected by the race/ethnicity and/or perceived gender role (in)congruity of the instructor; and do the types of qualitative student evaluation questions asked have different effects on students' subjective evaluations of their instructors? In the next section, I outline two theories of congruity, Role Congruity Theory (RCT) and Status Incongruity Hypothesis (SIH), as well as previous research on student evaluations of instruction (SEIs).

LITERATURE REVIEW

Role Theory and Social Role Theory are the foundational theories from which Role Congruity Theory (RCT) and Status Incongruity Hypothesis (SIH) were created. Role Theory considers how the roles people occupy and the expectations of those roles affects their own and others' behaviors, attitudes, and values (Jacobs 2018b). Roles happen at different levels; *specific roles* which occur in particular contexts and *diffuse roles* which occur in most contexts (Diekman and Schneider 2010b). For example, a person's occupation is a specific role because it occurs in the context of their workplace while their gender role is a diffuse role because it is salient in most social situations. The demands of specific roles can affect how a person operates in their diffuse roles and vice versa (Eagly et al. 2000).

Social Role Theory focuses on the effects of role expectations within the social structure and how a person is internally and externally affected when they occupy multiple roles, especially when the expectations of those roles are contradictory (Eagly and Karau 2002). Social Role Theory proposes that differences in the distribution of men and women into different social roles leads to differences in the observed behaviors and personalities of men and women (Eagly

et al. 2000; Koenig and Eagly 2014). Due to women being overrepresented in some specific roles and men being overrepresented in others, group stereotypes are formed such that it is assumed that it is "natural" for each gender to be in their corresponding role (Koenig and Eagly 2014). For example, women are historically overrepresented in caretaking roles such as child-rearing and home making and thus have come to be associated with being communal and caring (Eagly et al. 2000; Koenig and Eagly 2014; Wood and Eagly 2002). Men, on the other hand, have been historically overrepresented in the paid labor force and thus came to be associated with being agentic and dominant (Eagly et al. 2000; Koenig and Eagly 2014; Wood and Eagly 2002).

Role Congruity Theory

As described in the previous three chapters, Role Congruity Theory (RCT) is a subset of Role Theory and Social Role Theory. In brief, RCT posits that when a woman occupies a specific role that is incongruent with their gender role they may face two types of prejudice with respect to evaluations of their leadership: 1) less positive evaluations of their *potential* leadership abilities and 2) less positive evaluations of their *actual* leadership abilities (Diekman and Eagly 2008; Eagly et al. 2000). When women enter into masculine agentic domains and especially when they take on masculine leadership roles, they may be punished through negative evaluations or other sanctions such as not being considered for leadership roles and not being taken seriously when they are in leadership roles especially in traditionally masculine domains because of the perceived gender role incongruity (Eagly and Karau 2002; Eagly et al. 2000; Heilman 2012b). According to RCT, women can be seen as acceptable leaders which is a typically masculine role if they do so within a communal, feminine context such as when dealing with children/family, helping the poor, and/or working towards peace (Eagly and Karau 2002). However, when women are leaders in non-communal contexts they doubly violate feminine

gender norms through being a leader in a masculine domain and may elicit even stronger backlash reactions due to their participation in a masculine role (leader) in a masculine (non-communal) context (Eagly and Karau 2002).

Status Incongruity Hypothesis

Rudman et al. (2011) and Brescoll et al. (2018) argue that there are limitations to RCT. They argue that RCT: (1) only accounts for the experiences of women in leadership roles, (2) does not specify which aspects of gender roles cause backlash reactions, and (3) does not identify the motivations for penalizing targets who are role incongruent (Brescoll et al. 2018; Rudman et al. 2011)**.** Therefore, they propose Status Incongruity Hypothesis to mitigate the problems they identify with RCT.

SIH posits that backlash reactions occur when a person deviates from their prescribed gender norms one those deviations are perceived as a threat to the gender hierarchy (Moss-Racusin, Phelan, and Rudman 2010; Rudman et al. 2011). SIH argues that the motivation for backlash reactions is a defense of the existing gender hierarchy in which women are subordinate and less powerful than men. Thus, people who violate gender stereotypes that are most tied to the established gender hierarchy are more likely to receive backlash reactions (Moss-Racusin et al. 2010). Backlash can be both negative evaluations such as those described in RCT as well as negative emotional reactions such as feelings of disgust and visceral moral outrage due to the perceived threat the incongruity is to the traditional gender hierarchy (Moss-Racusin et al. 2010). For example, women who are in agentic leadership roles tend to experience backlash such as moral outrage and harsher evaluations than their men peers because the women are threating the traditional gender hierarchy by violating gender stereotypes of feminine communality and masculine agency through entering a masculine domain (Brescoll et al. 2018; Moss-Racusin et

al. 2010). Women in agentic masculine leadership roles are seen as especially threatening because they are taking a position of power which is a masculine role in a masculine domain thus positioning themselves hierarchically above other men in the domain when women are traditionally proscribed to subordinate roles especially in masculine domains (Moss-Racusin et al. 2010; Rudman et al. 2011).

Limitations of Current Congruity Theories

RCT and SIH both provide explanations as to why a person who is in a role that is incongruent with their gender may receive sanctions such as negative evaluations from others. Thus, RCT and SIH are useful frameworks for examining how target gender may affect the completion of subjective evaluations. They do, however, make different claims regarding the cause of backlash reactions to gender role incongruity. RCT posits that perceived role incongruity leads to sanctions while SIH posits that backlash is caused by a specific visceral reaction and the need to defend the gender hierarchy (Diekman and Eagly 2008; Rudman et al. 2011). Role congruity research needs to further tease out the causality of backlash responses in order to determine which if either of these two theoretical frameworks is more accurate.

Study Context: Congruity Theories and Student Evaluations of Instruction

The previous two quantitative studies (Chapters 2 and 3) of student evaluations of instructors found that unique combinations of instructor identities do effect student evaluations. The results of Study 1 (Chapter 2) showed that quantitative student evaluations are affected by the gender and gender-dominance of the discipline of the instructor being evaluated. The results of Study 2 (Chapter 3) added another layer to these results by showing that differences in quantitative student evaluation scores is further affected by the race/ethnicity of the instructor being evaluated. While these results are compelling, on their own they cannot speak to the

reasons *why* students perceive and evaluate their instructors of different identities and levels of role (in)congruity differently. Quantitative studies of student evaluations can only show that there are differences in evaluation scores and cannot assess the potential causal mechanisms for these differences. Furthermore, RCT and SIH provide different possible causal explanations as to why role incongruent persons receive more negative evaluations than their role congruent counterparts. Thus, quantitative outcomes for each theory will present the same but the underlying causal mechanism for the outcomes are different. In order to determine the validity of the possible causes for evaluation differences in RCT and SIH, an analysis of the *qualitative open-response* student evaluation questions is required. Qualitative analyses of students open response answers may help to discern which theory's propositions more accurately predict the causes of the differences in students' evaluations of their instructors through examining patterns in language especially with respect to backlash and threats to the gender hierarchy.

RCT predicts that role incongruent women instructors may receive backlash due to their defiance of traditional gender roles (Diekman and Eagly 2008; Eagly and Karau 2002; Heilman 2012b). Backlash against women teaching in traditionally masculine domains in the case of SEIs would lead to lower quantitative scores and negative qualitative comments. RCT does posit that some of the backlash women experience from occupying a masculine role can be avoided by over-emphasizing feminine characteristics in other ways (Heilman 2012b). With respect to student evaluations, this may mean that women why defy gender norms by teaching in traditionally man-dominated disciplines such as science and engineering may be able to mitigate some of the expected backlash through overdoing feminine aspects of their roles such as nurturing, caring, and helping students both during and outside of class time. Quantitative SEI measures do not ask questions about the additional tasks role incongruent instructors and

especially role incongruent women instructors may be doing to combat backlash on their student evaluations.

According to SIH, students may negatively evaluate role incongruent faculty, and role incongruent women in particular, specifically because they are perceived as a threat to the gender or other status hierarchy (Rudman et al. 2011). Empirical evidence from multiple studies (Brescoll et al. 2018; Moss-Racusin et al. 2010; Rudman et al. 2011) has found support for SIH and the notion that backlash to role incongruity occurs when the accepted status quo is challenged. Women who teach in STEM may be seen as a threat to the gender hierarchy which may result in moral outrage on the part of the student which is then communicated through the open-ended SEI questions as the close-ended SEI questions do not offer a particular opportunity for students to partake in backlash. Furthermore, if a woman teaching in STEM is particularly verbose or domineering, she may be seen as even more of a threat to the gender hierarchy thus resulting in even more negative comments from students that are driven by moral outrage (Brescoll et al. 2018).

Through an examination of the qualitative open-response questions on student evaluation forms, the claims of RCT and SIH can be tested and the potential causes of students' unconscious biases which lead to differences in student evaluation scores can be determined. Furthermore, more insights into the potential extra burdens facing some groups of faculty (women and racially/ethnically minoritized faculty) and not others (White men) may be discernable through reading the comments from students about their experiences with the instructor. It is important to gain a better understanding of how the written comments from students may vary based on the characteristics of the instructor because the written comments from students can have a profound impact on how teachers view themselves, their teaching style,

and their ability to lead a classroom. Yet, it is often the case that faculty are only presented with how they compare to their colleagues on the quantitative student evaluation measures. Therefore, it is crucial to understand how these less easily quantifiable and comparable but just as important student evaluation measures may be systematically punishing some groups of instructors while systematically praising others.

Study Context: Previous Qualitative Research on Student Evaluations of Instruction

RCT and SIH both provide explanations for why a person who deviates from traditional gender roles may receive sanctions such as negative evaluations. Previous qualitative studies of student evaluations have not specifically examined the predictions of these two theories, but they have found that there are clear gender differences in the ways in which students talk about their instructors. According to an analysis of comments on RateMyProfessor.com, students tend to describe their men professors using words like brilliant, intelligent, and expert while women tend to be described as mean, nice, rude, demanding, and crazy (McMurtrie 2019). In another study, students were asked to provide adjectives to describe the best and worst teacher they ever had (Sprague and Massoni 2005). The results of this study showed that while there is overlap in how students talk about their men and women instructors, there are still clear gendered differences (Sprague and Massoni 2005). Students tended to describe their best men teachers as funny while describing their best women teachers as caring and nurturing all of which are positive comments, however the comments about women instructors are tied to traditionally feminine communal gender role expectations (Sprague and Massoni 2005). The worst men teachers were described as boring and self-centered while the worst women teachers as rigid, mean, and unfair which are typical comments used as backlash against women who act agentically (Sprague and Massoni 2005). These are clearly gendered patterns of language that students use when they are allowed

to openly evaluate their instructors and especially women instructors who get either positive communal feedback or negative backlash.

Previous RCT and SIH research suggests that backlash reactions may not be universal for all racial groups as students may not have the same gendered expectations for Black women teaching in STEM as they do for White women (Biernat and Sesko 2013; Livingston et al. 2012). These different race and gender expectations may potentially result in less comments driven by moral outrage written about Black women due to their gender role incongruity than their White women peers (Livingston et al. 2012). However, student evaluation research consistently finds that racially/ethnically minoritized instructors tend to receive lower quantitative evaluation scores and less positive qualitative comments than White instructors (Aruguete et al. 2017; Bavishi et al. 2010; Chávez and Mitchell 2020; Reid 2010; Smith and Hawkins 2011). Student evaluations of instructors based on their gender and race/ethnicity may be further affected by their temperament and teaching style (Anderson and Smith 2005). An experiment in which students evaluated hypothetical instructors based on their course syllabi found that students rated White women instructors with a strict teaching style as warmer than Latinx women instructors with a strict teaching style (Anderson and Smith 2005). The opposite pattern emerged for instructors with a lenient teaching style—students rated Latinx women instructors as warmer than White women instructors (Anderson and Smith 2005).

Bavishi, Madera, and Hebl (2010) conducted an experiment in which students ranked instructors based on an examination of their CV which varied by gender, race, and academic discipline (science or humanities). While they did not find any gender main effects, their results indicate that students perceived White instructors as having more interpersonal skills than their Black and Asian counterparts and that science professors are perceived as more competent and

legitimate than humanities professors (Bavishi et al. 2010). Furthermore, there were interaction effects between instructor race and discipline such that African American humanities professors were evaluated as less legitimate than White science professors (Bavishi et al. 2010). Additionally, female professors in the humanities were rated as less competent than male professors in the sciences and African American female professors were shown to have the lowest ratings of all groups on measures of competence, interpersonal skills, and legitimacy (Bavishi et al. 2010). Their results indicate that race, much like gender and in combination with gender and discipline, may also affect students' expectations for the interpersonal skills of their instructors and their subsequent evaluations of their instructors. Taking both the anecdotal and empirical evidence together, the pattern of differences in qualitative evaluation comments indicates that students tend to qualitatively evaluate their men and women, White and racially/ethnically minoritized professors differently, which may be due to different expectations for people of different social statuses in the role of professor.

These nuanced differences in student evaluations based on instructors' gender, race/ethnicity, discipline, and other factors such as temperament may not emerge in studies of only quantitative student evaluation measures. An examination of qualitative student evaluation questions may better highlight the differences in students' perceptions, evaluations, and expectations of their instructors based on their gender, race/ethnicity, and discipline as well as test the claims of Role Congruity Theory and Status Incongruity Hypothesis. In this study, over 1,400 responses to open-response student evaluation questions are coded across groups of instructors of different genders, race/ethnicities, and gender-dominance of the discipline in which they teach in order to determine the potential causes for differences in students' evaluations of their instruction.

DATA

Student evaluation of instructor data as well as human resources data from a large research-intensive land-grant university in the Appalachian region of the United States were utilized in this study. See Chapter 1 for a complete description of the data cleaning, merging, and variable creation process. The data includes responses from five fall and spring academic semesters between fall 2016 to fall 2018. The questions on the student evaluation forms from fall 2016 and spring 2017 are the same but between spring 2017 and fall 2017 there were significant changes made to the student evaluation forms. These changes include completely new qualitative questions. Thus, the qualitative data coding was done in two parts: old questions and new questions. The "old questions" which were asked during fall 2016 and spring 2017 simply provide two open-response textboxes for students with the prompts (1) Comments on Course (Course) and (2) Comments on Instructor (Instructor). The "new questions" which were asked from fall 2017 to fall 2018 provided students with open-response textboxes to respond to the questions (1) What helped you learn in this course? (Helped Learn) and (2) What recommendations do you have for change? (Change). The number of responses per question can be seen in Table 32. The shift from broad statements asking for comments to more pointed questions about specific topics related to the course may lead to very different qualitative responses from students. The change in questions shifted the focus from instructors and courses generally to two specific areas of course feedback. This may lead to less comments on the instructor themselves and more comments on the course materials and teaching mechanisms. The coding from each time period will be compared in the analyses that follow in order to tease out any differences caused by the change in questions. These comparative analyses will determine

the effect of the qualitative question changes on student evaluation responses and how it relates to the level of role congruity of instructors.

Table 32: Qualitative Questions and Response Counts

Question Name in Text	Student Evaluation Question	Response Count	Semesters
Course	Qualitative comments "Comments on Course:"	23,129	Fall 2016, Spring 2017
Instructor	Qualitative comments "Comments on Instructor:"	33,090	Fall 2016, Spring 2017
Helped Learn	Qualitative question "What helped you learn in this course?"	68,018	Fall 2017, Spring 2018, Fall 2018
Change	Qualitative question "What recommendations do you have for change?"	55,037	Fall 2017, Spring 2018, Fall 2018

METHODS AND ANALYSES

The data were sorted by faculty gender, race/ethnicity, and discipline. Gender was separated into two categories, man and woman, with all others removed from the sample. Race/ethnicity was separated into two categories, White and racially/ethnically minoritized, with all those with race not reported removed from the sample. Race/ethnicity was collapsed in this way due to small sample sizes in the individual race/ethnicity groups. The disciplines of the instructor were separated into three categories, man-dominated, woman-dominated, and neutral. The full process by which disciplines were categorized is described in Chapter 1. In order to avoid interdependence between the responses for each question, responses for different instructors were used for each of the two questions in each of the two time periods. Thus, there are 12 instructor groups across two time periods and two questions in each time period for a

grand total of 48 instructor categories (12 groups x 2 periods x 2 questions = 48)[14]. The descriptions of the faculty groups can be found in Appendix 17.

Prior to sampling, blank and, when appropriate, "filler" (i.e. no comment, n/a) responses were eliminated so that the sampling only occurred on legitimate substantive responses. For the questions Helped Learn and Change, it was important to leave in comments such as "nothing" or "none" as the questions were specifically asking for what helped students learn and recommendations for change, respectively. In these cases, unlike the earlier two questions, it was important to capture these types of "filler" responses as they were perfectly reasonable answers to the newly worded questions. After removing filler responses for the question being sampled for, six faculty for each of the 48 categories were selected for a total of 288 unique faculty. The selected faculty were semi-randomly chosen. For discipline groups which were the same (e.g. man-dominated), responses for instructors teaching in the same or very similar fields were randomly selected. For example, for instructors in man-dominated disciplines, fields such as engineering, physics, and finance were consistently sampled from across the gender, race/ethnicity, and question groups of faculty. Whenever possible, the level of the course responses were selected from was matched for each field across instructor groups. For example, if a 100-level engineering course was sampled from for White men in man-dominated disciplines, a similar level and discipline course was sampled from for the other instructor categories for each question.

From the six faculty selected for each of the 48 faculty groups, five random responses from one course were selected[15]. Responses were chosen from the same course for each faculty

[14] Another way to think about this is that there are 4 questions and I sampled from a different instructor from each category for each question (4 questions x 12 instructor categories).

[15] There was 1 faculty category in which there were not enough unique responses for one class from six different instructors that being racially/ethnically minoritized women in neutral disciplines on the older SEI forms. In this

because different courses may lead to different responses even for the same instructor. For example, higher level courses tend to be taken by students who are more invested in the topic area than introductory level courses and therefore the qualitative feedback of the two courses may be very different. Thus, by only selecting responses from one course for each instructor, the randomly selected responses provide a more complete evaluation of that particular course taught by that particular instructor. Thus, 30 responses were semi-randomly selected for the 48 instructor groups for a total of 1,430 responses.[16] The courses sampled from were from a variety of course levels within each instructor group. Lower-level (100 and 200) courses and higher-level (300 and 400) courses were sampled from in each instructor group whenever possible such that responses from a variety of course levels were chosen for each instructor group. Thus, every attempt was made to try to mitigate any potential biases from over selecting responses from lower- or higher-level courses.

Each response was coded with six themes in mind: positive professional, negative professional, positive personal, negative personal, positive course, and negative course. Comments were marked positive and professional when they commented positively on the intelligence/expertise of the instructor, their teaching style, or their general ability to teach the course. Comments that referred to the professor as a "good professor" or made specific reference to teaching style were coded as positive professional comments. Comments were marked negative and professional when they commented negatively on the intelligence/expertise of the instructor, their teaching style, or their generally ability to teach the course. For example,

case, five samples were taken from two different classes for one instructor. This was avoided whenever possible in order to mitigate issues of interdependence as much as possible.

[16] There should have been 1,440 responses, however, due to a lack of unique responses for the older SEI forms for racially/ethnically minoritized women in neutral disciplines, only five unique classes were sampled from for each question thus giving a total of 1,430 qualitative responses in the sample.

comments that indicated they had to teach the material to themselves due to the class structure were coded as a negative professional comment.

Comments were marked positive and personal when they commented positively on the nature of the professor rather than their ability to teach the course. Comments that referred to the professor as "understanding", "exciting", or a "good person" were coded as positive personal comments. Comments were marked negative and personal when they commented negatively on the nature of the professor rather than their ability to teach the course. Comments that referred to the instructor as "mean", "rude", or "not personable" were coded as negative personal comments.

Comments were marked positive course when they commented positively on the course and/or the materials in the course rather than about the instructor or their teaching style. Comments that mentioned liking the assignments, finding study guides helpful, or enjoying the labs were coded as positive course. Finally, comments were marked negative course when they commented negatively on the course and/or the materials in the course rather than about the instructor or their teaching style. Comments that mentioned the work being hard, lectures being boring, or the course itself lacking structure were coded as negative course. While the two course codes do reflect on the ability of the instructor to teach and do their profession, the course codes were made distinct from the positive/negative professional codes because they focused specifically on the materials or class and not on the instructor's intelligence, expertise, or professionalism.

A sub-sample of responses was cross coded by three people and the codes were compared for consistency. Any inconsistencies in coding were addressed to mitigate issues of coder reliability. Table 33 contains an example response for each code where each bolded portion represents a separate idea that was counted towards the code count for that post. For example, in

the positive professional example response, there are two sections bolded so that comment would have been given a positive professional score of 2.

Table 33: Example Student Evaluation Response for Each Code Type

Code	Student Response
Positive Professional	**Smart** professor that **knows what he is talking about**
Negative Professional	He was **not the best professor** and he **did not really help me learn** the material better. He **read everything off the powerpoint.**
Positive Personal	Dr. Mxxxxx is a **caring compassionate** and wise professor. Her class has turned me into a better person.
Negative Personal	**Rude**…He also made **inappropriate comments** like "**who's your daddy**"
Positive Course	**Good course** especially with **developing communication skills**
Negative Course	**Very hard class!** it sucks that it **wasn't stuff about what i want to do.**

Note: Qualitative responses were examined exactly as they were written and were not edited for spelling or grammar.

Theoretical Predictions

In combination, the course, professional, and personal codes paint a very descriptive picture of how students are thinking about and evaluating their instructors as people, as professionals, and the course content they utilize. Furthermore, these specific code types allow for an assessment of the potential causal mechanisms proposed by RCT and SIH. Firstly, Role Congruity Theory posits that backlash is a response to women who enter into leadership roles especially in masculine domains whereas women can be seen as acceptable leaders when in feminine domains (Eagly and Karau 2002; Eagly et al. 2000; Heilman 2012b). Thus, the professional comments are especially relevant as harsher comments about the professionalism of women instructors in man-dominated disciplines would indicate support for RCT. Secondly, Status Incongruity Theory posits that backlash is caused by a specific visceral reaction to violations of traditional gender roles which are perceived to threaten the gender hierarchy and are expressed through disgust and moral outrage (Moss-Racusin et al. 2010; Rudman et al. 2011).

Thus, if there are harsher personal comments which express feelings of rage towards any instructor who teaches outside of their traditional gender domain that would indicate support or SIH. Finally, the course codes are not as directly related to the propositions of RCT and SIH but preliminary coding suggested the necessity for their inclusion due to the volume of comments related to aspects of the course which are not directly related to the instructor. Additionally, if there are more negative course comments for perceived role incongruent instructors than perceived role congruent instructors, that would suggest that perceptions of gender role congruity may affect students' evaluations of courses even if they are not directed specifically towards the instructors.

If comments in general tend to penalize women instructors in man-dominated disciplines while not doing the same for men instructors in woman-dominated disciplines and tend to do so more on the professional code than any other code, that would indicate that RCT may be the more correct theory. Instead, if comments tend to punish any violations of traditional gender roles and especially through expressions of disgust and outrage more so on the personal code than any other code, that would indicate that SIH is the more correct theory. If comments tend to punish all violations of perceived gender role incongruity but do not do so through expressions of disgust and outrage on any of the three codes, that would suggest that both RCT and SIH have some merit and therefore scholars should take the propositions of both theories into consideration when studying the effects of perceptions of gender role congruity on perceptions and evaluations.

Analytical Process

When coding, each separate instance of each theme was demarcated in each response. A score was tallied for each category for each theme. A total overall score for each response was

then tallied with positive comments counting positively and negative comments counting negatively towards the overall score. For instance, if a response had one positive professional comment and one negative personal comment the overall score for the response would be zero. Overall scores were calculated to best represent the overall mood of the response and whether that was positive, negative, or neutral. Net scores for each response were also calculated in which all responses were simply added together, regardless of if they were negative or positive responses. The net scores were calculated to show which instructor groups get the most comments, regardless of the sentiment. These overall and net scores were examined to get a general sense of the overall feel of the responses and to note any major differences between faculty categories. The scores for each code category provided more details about how students were talking about each instructor group. While the net scores obfuscate the direction of the sentiment, positive or negative, examining the net scores in combination with the individual code scores paints a vivid picture of which instructors are talked about the most (net scores) and how they were talked about (individual code scores).

The responses were also examined qualitatively to determine the major themes of the qualitative responses. Due to the large amount of data, the qualitative thematic coding and analyses began with sentiment analyses and word clouds to get a sense of the feel and major themes of the responses by instructor group. These two analyses were conducted in R statistical software for each instructor group. Sentiment analyses of the comments for each instructor group were completed to get a general sense of the emotional feel of the responses. Sentiment analysis use natural language processing techniques to identify the emotional tone of the text. The sentiment analyses used a predetermined schema to classify responses based on ten sentiment categories: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative, and positive.

A bar chart of the count of responses classified in each sentiment was created for each instructor group. While sentiment analyses are not perfect as tone and sarcasm can obfuscate the results, this process did provide a general sense of the overall tone and emotion of the responses in an efficient manner.

Word clouds were also created to get a general sense of the main themes for each instructor group. Word clouds identify the most used words in a dataset, in this case all responses for the twelve instructor groups. Words that are used more often appear larger and darker in the word cloud while words that are used less often appear smaller and lighter. The top thirty most used words were included in the word cloud for each instructor group.

Finally, for each of the 48 instructor question groups, summaries of each of the six code themes were written. These summaries were then combined by the twelve instructor groups for comparison. The summaries were examined at the instructor level (12 groups) rather than the instructor question level (48 groups) to more easily determine if there were major differences in the way in which students talk about their instructors as influenced by the gender, gender role (in)congruity, and/or race/ethnicity of the instructor being evaluated. The qualitative theme summaries as well as the quantitative descriptive analyses are presented below.

RESULTS AND DISCUSSION

The quantitative code counts are presented first followed by the qualitative themes. Within each set of results, the implications of the instructors' gender, gender role congruity, and race/ethnicity are discussed.

Quantitative Description

Average scores for each code and time period as well as the overall score and net score averages can be found in Table 34. The average quantitative code score for each instructor group

and question can be found in Appendix 18. The overall score averages for both the old (Course and Instructor) and new (Helped Learn and Change) questions were positive with the old questions being even more positive than the new questions. This indicates that while students do write negative comments about the instructor and the course, a majority of comments are positive in some way.

Table 34: Average Score for Each Coding Category and Time Period

	Positive Professional	Negative Professional	Positive Personal	Negative Personal	Positive Course	Negative Course	Overall Score	Net Score
Avg Of Old	**0.74**	**-0.31**	**0.23**	**-0.07**	0.49	-0.45	0.62	2.30
Avg Of New	0.29	-0.18	0.07	-0.01	**0.64**	**-0.53**	0.26	1.74
Avg Of All	0.51	-0.24	0.15	-0.04	0.56	-0.49	0.44	2.02

The old questions, Course and Instructor, led to more extreme responses in the professional and personal codes while the new questions, Helped Learn and Change, led to more extreme responses on the course codes. So, in effect, when the institution changed the questions on the SEI forms, they shifted the focus of students' responses from talking directly about faculty to talking about the substance of the class itself. Comments on the old (Course and Instructor) questions tend to say something about the instructor's intelligence, their personality, and/or their ability to teach the course well while comments on the new (Helped Learn and Change) questions tend to say more about the assignments, readings, and/or other course materials.

The switch from talking about the instructor to talking about the course is good in some ways as there are less comments made about the instructor as a person, but it also means that there is less feedback about the actual instructors, good or bad. Additionally, some of the course comments are about the scheduling of the course or the general topic (e.g. History) which are outside of the control of the individual instructor. For example, several students commented

about time a course occurred. It is not necessarily the individual instructor's fault that, for example, class is held in the mornings. However, many of the course-related comments are likely within the instructor's control such as the textbook, assignments, and the general organization of the course. Thus, there are pros and cons to the switch away from comments being about the instructor to mostly comments about the course content which may or may not be under the purview of the instructor.

There were the least personal comments of either type, positive or negative, across all questions as compared to any other category. Table 35 describes the number of instructor categories that had zero comments in each code. Of the 48 instructor question categories, 21 (43.75%) did not have any positive personal comments and 32 (66.67%) did not have any negative personal comments. On the other hand, almost every Instructor category had at least some comments about the course in either direction and most had comments either positive or negative about the instructor professionally.

Table 35: Number of Categories with an Average Score of Zero Indicating No Responses

	Positive Professional	Negative Professional	Positive Personal	Negative Personal	Positive Course	Negative Course	Overall Score	Net Score
All Questions	6	11	21	32	2	0	0	0
Instructor	5	4	11	11	0	0	0	0
Course	0	0	0	3	2	0	0	0
Helped Learn	0	6	0	10	0	0	0	0
Change	1	1	10	8	0	0	0	0

There were only 2 of the 48 categories which did not have positive course comments and there were no categories that did not have any negative comments about the course. These results indicate that on the whole, there are less personal comments about the instructor than there are comments about the course. According to Table 36, when there were positive personal comments, White men neutral (neutral), racially/ethnically minoritized women woman-

dominated (congruent), racially/ethnically minoritized men neutral (neutral), White women man-dominated (incongruent), and White women neutral (neutral) had highest average scores. More instructor groups from neutral disciplines elicited positive personal comments than instructor groups in woman- or man-dominated disciplines. This indicates that with respect to positive personal comments, being in a neutral discipline may be an advantage. With respect to negative personal comments, the most comments were made about White men woman-dominated (incongruent), racially/ethnically minoritized men woman-dominated (incongruent), racially/ethnically minoritized women man-dominated (incongruent), White men man-dominated (congruent), and racially/ethnically minoritized women woman-dominated (congruent). Thus, negative personal comments are split between instructors who teach in incongruent and congruent disciplines, though the most extreme negative scores do belong to instructors who teach in gender role incongruent fields. Furthermore, there is a mix of both role incongruent men and women who tend to receive the most negative personal comments indicating that the driving force may be less about gender itself and more about the violation of gender roles. Thus, negative feedback is less about *who* violates (men or women) but rather more about a violation of gender norms in general. Both men and women who violate gender norms were punished through receiving negative personal comments. More notably, both role incongruent racially/ethnically minoritized men and women received the most negative personal comments indicating that race/ethnicity may further drive negative personal comments especially when instructors teach in role incongruent disciplines. The only instructor group which elicited one of the highest volumes of both positive personal and positive negative comments was racially/ethnically minoritized women in woman-dominated disciplines who are therefore role congruent with the discipline in which they teach. Again, this indicates that race/ethnicity may

affect students making comments about the instructor as a person more so than gender role congruity.

Table 36: Average Quantitative Code Score by Instructor Group

Instructor Group	Perceived Level of Congruity	Positive Professional	Negative Professional	Positive Personal	Negative Personal	Positive Course	Negative Course	Average of Codes	Average Overall Score	Average Net Score
White Man Man-Dominated	Congruent	0.48	-0.28	0.12	-0.07	0.59	-0.66	0.03	0.18	2.23
White Man Woman-Dominated	Incongruent	0.43	-0.15	0.13	-0.08	0.64	-0.38	0.10	0.60	1.80
White Man Neutral	Neutral	0.62	-0.09	0.22	-0.01	0.52	-0.48	0.13	0.77	1.93
White Woman Man-Dominated	Incongruent	0.64	-0.32	0.18	-0.03	0.48	-0.37	0.10	0.58	2.01
White Woman Woman-Dominated	Congruent	0.52	-0.14	0.09	0.00	0.72	-0.57	0.10	0.61	2.03
White Woman Neutral	Neutral	0.44	-0.13	0.16	-0.01	0.57	-0.43	0.10	0.60	1.73
Racially/ethnically Minoritized Man Man-Dominated	Congruent	0.45	-0.23	0.13	-0.03	0.60	-0.58	0.06	0.35	2.02
Racially/ethnically Minoritized Man Woman-Dominated	Incongruent	0.43	-0.30	0.13	-0.08	0.68	-0.43	0.07	0.43	2.04
Racially/ethnically Minoritized Man Neutral	Neutral	0.52	-0.33	0.18	-0.04	0.45	-0.52	0.04	0.27	2.04
Racially/ethnically Minoritized Woman Man-Dominated	Incongruent	0.54	-0.45	0.13	-0.09	0.58	-0.43	0.05	0.29	2.23
Racially/ethnically Minoritized Woman Woman-Dominated	Congruent	0.57	-0.16	0.20	-0.07	0.47	-0.58	0.07	0.35	2.15
Racially/ethnically Minoritized Woman Neutral	Neutral	0.53	-0.34	0.14	-0.04	0.48	-0.50	0.04	0.28	2.05

The instructor groups with the highest scores on the positive professional code were White women man-dominated (incongruent), White men neutral (neutral), racially/ethnically minoritized women woman-dominated (congruent), racially/ethnically minoritized women man-dominated (incongruent), and racially/ethnically minoritized women neutral (neutral). With respect to perceived congruity, the groups with the highest positive professional scores are mixed with two incongruent, two neutral, and one congruent group all receiving the highest scores. Women and especially racially/ethnically minoritized women were more likely to receive positive professional comments with White perceived incongruent and racially/ethnically minoritized women of all perceived congruity levels receiving the highest scores. While this may indicate that racially/ethnically minoritized women are at an advantage with respect to receiving positive professional comments on student evaluations, examining the results of the negative professional comments challenges this idea.

For the negative professional code, the groups with the highest scores were racially/ethnically minoritized women man-dominated (role incongruent), racially/ethnically minoritized women neutral (neutral), racially/ethnically minoritized men neutral (neutral), White women man-dominated (incongruent), and racially/ethnically minoritized men woman-dominated (incongruent). Thus, while racially/ethnically minoritized women in man-dominated and neutral disciplines receive high amounts of positive professional comments they also receive high amounts of negative professional comments. Additionally, racially/ethnically minoritized men in women-dominated and neutral disciplines also receive some of the highest amounts of negative professional comments. Taken together, these results indicate that when racially/ethnically minoritized instructors teach in perceived role incongruent or neutral disciplines, they are more likely to be negatively perceived and evaluated with respect to

professional aspects. This indicates that role congruity may lead to more negative comments of the professional aspects of instructors when they are racially/ethnically minoritized than when they are White.

Finally, the instructor groups with the highest scores on the positive course codes were White women woman-dominated (congruent), racially/ethnically minoritized men woman-dominated (incongruent), White men women-dominated (incongruent), racially/ethnically minoritized men man-dominated (congruent), White men man-dominated (congruent), and racially/ethnically minoritized women man-dominated (incongruent). Men, whether in perceived role congruent or role incongruent disciplines, tend to receive the highest amounts of positive course comments. For women, perceived gender role congruity led to more positive course comments when the instructor was White but perceived gender role incongruity led to more positive course comments when the instructor was racially/ethnically minoritized.

The groups with the highest amounts of negative course comments were White men man-dominated (congruent), racially/ethnically minoritized men man-dominated (congruent), racially/ethnically minoritized women woman-dominated (congruent), White women woman-dominated (congruent), racially/ethnically minoritized men neutral (neutral), and racially/ethnically minoritized women neutral (neutral). The three congruent groups who received the most negative course comments (White men, White women, and racially/ethnically minoritized men) also received the most positive course comments while the two perceived neutral groups (racially/ethnically minoritized men and racially/ethnically minoritized women) only received a high amount of negative course comments. This indicates that perceived role congruity may lead to more balanced evaluations of course materials, students express liking one thing and disliking another, whereas role neutrality leads to mostly negative course comments

especially for racially/ethnically minoritized instructors. The only perceived congruent group to receive a high amount of negative but not positive course comments was racially/ethnically minoritized women which indicates that while perceived role congruity may help to balance the evaluations of some instructor groups (White men, White women, and racially/ethnically minoritized men) the balancing effects are not universal.

Furthermore, when racially/ethnically minoritized instructors teach in disciplines with which they are perceived to be role incongruent they tend to receive more positive course comments while role congruent and neutral racially/ethnically minoritized instructors tend to receive more negative course comments. Racially/ethnically minoritized men in man-dominated disciplines (congruent) tend to also receive a higher amount of positive course comments than other groups. The high amount of course comments for all racially/ethnically minoritized instructors indicates that race/ethnicity may be a driving force behind students perceptions and evaluations of courses when completing their student evaluations of instruction. It may be that students have stronger opinions in general about racially/ethnically minoritized instructors but are cautious of making comments about them as a person, either professional or personal, positive or negative, for fear of appearing to be racially motivated. Instead, they may opt to express their enthusiasm and grievances through making more comments about the course and the course materials which may pose less of a potential personal threat to the student by mitigating the risk of appearing racially/ethnically intolerant. It may also be that the anonymous nature of SEIs may also take away students' fears of appearing to be racially motivated in their negative feedback thus leading to more negative responses for racially/ethnically minoritized instructors than White instructors. Thus, students feel comfortable blatantly expressing negative

sentiment about racially/ethnically minoritized instructors because they are told that the instructor will not know who is giving the particular feedback.

Table 37 contains the categories with the most extreme scores for each code. Of these categories, three of the five are from the Instructor question on the old SEI forms. This question asked for "comments on the instructor". The faculty groups which produced the most extreme scores on this question were White men in perceived neutral disciplines (neutral) who were most extreme on the positive professional, positive personal, overall, and average of the separate category scores. Also on the Instructor question from the old SEI forms, racially/ethnically minoritized men in perceived neutral disciplines (neutral) were most extreme in the positive personal comments while racially/ethnically minoritized women in man-dominated disciplines (incongruent) were most extreme on the negative professional and negative personal comments. One of each of the questions on the new SEI forms represent the other two categories with the most extreme responses, those being Help Learn and Change. For the question Helped Learn which asks "what helped you learn in this course", White women in woman-dominated disciplines (congruent) elicited the most extreme response in the positive course comments. For the Change question which asked about recommended changes, White men in man-dominated disciplines (congruent) elicited the most extreme average score in the negative course comments.

Table 37: Categories with the Most Extreme Score for Each Category

Form	Old	Old	Old	New	New
Instructor Group	White Man Neutral	Racially/ethnically Minoritized Man Neutral	Racially/ethnically Minoritized Woman Man-Dominated	White Woman Woman-Dominated	White Man Man-Dominated
Level of Congruity	Neutral	Neutral	Incongruent	Congruent	Congruent
Question	Instructor	Instructor	Instructor	Helped Learn	Change
Positive Professional	**1.87**	1.40	1.50	0.57	0.13
Negative Professional	-0.13	-0.53	**-1.30**	0.00	-0.47
Positive Personal	**0.67**	**0.67**	0.40	0.03	0.00
Negative Personal	-0.03	-0.17	**-0.37**	0.00	-0.10
Positive Course	0.00	0.23	0.07	**1.43**	0.10
Negative Course	-0.03	-0.33	-0.37	-0.17	**-1.93**
Overall Score	**2.33**	1.30	-0.07	1.83	-2.23
Average of Separate Categories	**0.39**	0.21	-0.01	0.31	-0.38
Net Score	2.73	3.33	**4.00**	2.20	2.73

These results are mixed with both men and women, White and racially/ethnically minoritized and all levels of perceived (in)congruity—neutral, incongruent, and congruent—instructors receiving the most extreme comments but most of the most extreme comments occurred when students were simply asked for "Comments on the Instructor". When perceived gender role congruity is also examined, men in neutral disciplines, and especially White men, tend to receive the highest amounts of positive comments, especially on the Instructor question. These findings indicate that when men do not adhere strictly to gender roles by working in a neutral space, they may still be perceived and evaluated positively.

Racially/ethnically minoritized women who teach in perceived gender role incongruent disciplines (man-dominated) tend to receive the highest amount of negative personal and professional comments on the Instructor question as well as the highest net amount of comments. This indicates that when racially/ethnically minoritized women in particular violate gender norms, students may provide more commentary in general and evaluate them more negatively with respect to their professional and personal attributes when asked directly about the instructor. No other perceived role incongruent group received a high amount of negative codes for any question indicating that students may more harshly evaluate racially/ethnically minoritized women in perceived gender role incongruent disciplines than perceived gender role incongruent racially/ethnically minoritized men, White men, or White women. Racially/ethnically minoritized women in role incongruent disciplines violate gender norms by being a leader in their classroom, by teaching in man-dominated disciplines, and by entering into a traditionally White dominated occupation. Thus, the greater number of negative evaluations received by racially/ethnically minoritized women may be caused by the unique intersection of these women's racial, gender, and gender role identities.

Finally, White perceived gender role congruent men and women received the most course comments with women being evaluated positively when asked "what helped you learn" (Helped Learn) and men being evaluated negatively when asked for "recommendations for change" (Change). White perceived gender role congruent women were more likely to be perceived as providing helpful materials while White gender role congruent men were more likely to be perceived as needing to improve some aspect of their course. Women and White women, in particular, are stereotypically perceived as being helpful and therefore it is not all that surprising that they received the highest positive course score on the Helped Learn question. Men and White men, in particular, may be seen as stuck in their ways and unwilling to change, thus leading to more negative course comments calling for change on the Change question. These two findings indicate that an instructor's gender and the fact that they are perceived to be gender role congruent may be subconsciously influencing their perception of not only the instructor but of the course and its materials especially when the instructor being evaluated is White.

Qualitative Themes

To gain a better understanding of how students talk about their instructors and how this varies by the gender, perceived gender role (in)congruity, and/or race/ethnicity of the instructor, more qualitative thematic processing was completed. For the qualitative thematic process, sentiment analyses (Appendix 19) and word clouds (Appendix 20) were examined first to get a general sense of the feel and main themes of the responses by instructor group. Then, summaries (Appendix 21) of all the comments for each code and instructor group were written and compared allowing for more nuanced themes to emerge from the data.

In examining the sentiment analyses by instructor group and perceived level of gender role (in)congruity, the bar charts echo the results from the quantitative analyses above in that the

highest sentiment for every instructor group is positive followed by trust. Students, for the most part, tend to write positive comments about their instructors regardless of gender, gender role (in)congruity, or race/ethnicity. Disgust did appear to be higher for perceived role incongruent White men, racially/ethnically minoritized men, and racially/ethnically minoritized women than role incongruent White women. This finding indicates that perceived violations of gender roles may lead to more perceptions of disgust for men and racially/ethnically minoritized women who violate gender roles than White women. This finding suggests that SIH may be correct in the proposition that men who are perceived to violate gender norms may face repercussions through expressions of moral outrage which in this case is specifically expressed as disgust. Otherwise, there are not obvious differences in the sentiment analyses of the twelve different instructor groups. All twelve of the bar charts follow a similar pattern with just slight differences between them.

Similarly, the word clouds indicate many of the same words are used most frequently regardless of the instructor group, though there are a few subtle differences in frequency of use worth noting. For perceived role congruent faculty, commonly used words included learn, material, help, test, understand, good, and great. For perceived role incongruent faculty, commonly used words included learn, understand, material, help, test, time, good, and great. For perceived role neutral faculty, commonly used words included learn, test, material, great, help, student, understand, assignment, and teacher. An instructor's name was also identified as a most used word for neutral racially/ethnically minoritized women instructors indicating that students used her name frequently enough in their comments that it was identified as being in the top 30 used words. This is the only such occurrence in any of the twelve word clouds and therefore it is likely an anomaly related to this particular instructor but it could be connected to how students

perceive and evaluate racially/ethnically minoritized women instructors in perceived neutral disciplines. Upon closer looking, this instructor's name was used by one student multiple times to describe them with positive personal and positive professional comments. Given that the usage is only by one student, this is unlikely indicative of a larger pattern of talking about role neutral racially/ethnically minoritized women but rather an anomalous comment by one student who particularly liked the instructor. While using instructors' names appears to be an anomaly as it is the only occurrence in the sample, it could occur again. While the sampling was randomly completed and therefore should be representative of the whole population of qualitative responses, it is still only a sample and therefore it cannot be definitively stated that this does not occur again. However, on the whole much like as is shown by the sentiment analyses and quantitative analyses above, the most frequently used words across all instructor groups are fairly positive in nature which reflects the larger trend of students being overwhelmingly positive in their evaluations of instruction.

Themes from Code Summaries

While sentiment analyses and word clouds helped to orient further thematic processing of the data, a closer examination of the responses by code theme was necessary to examine if the perceived gender role congruity of the instructor affects the ways in which students respond to open-ended student evaluation questions. All comments from each code for each instructor group were grouped and summarized (Appendix 21). Positive professional comments, while they do vary in quantity as discussed above, all tended to be pretty similar qualitatively across the instructor groups with many instructors being described as good, great, knowledgeable, and helpful. There were some other common themes among the posts which tended to vary by instructor gender, perceived gender role (in)congruity, and/or race/ethnicity.

Students often talked about grades in their evaluations whether that be the perceived fairness/unfairness of the instructor or the timeliness/lack thereof of grades being posted to the learning management system for the course. Women instructors tended to be described as unfair or harsh graders regardless of their perceived level of gender role congruity with White and racially/ethnically minoritized congruent, incongruent, and neutral women all being described this way. These findings indicate that women, regardless of the discipline they teach in and their race/ethnicity, are likely to be perceived as unfair graders therefore it may be that students view women as instructors in any discipline as being out of place with their gender role due to the authority and leadership associated with the role of instructor. Students expect all women, regardless of discipline, to be communal and compassionate and therefore easier graders. Thus, when women in any discipline assert their authority through grading students in any way that is not strictly positive, the woman instructor it likely to be perceived as being harsh due to the contrast with feminine gender norms no matter the discipline in which they teach.

Of men instructors, only racially/ethnically minoritized men teaching in woman-dominated disciplines (incongruent) and White men in neutral disciplines (neutral) were described as harsh graders. Furthermore, two men instructor groups were described as fair graders those being White men in women-dominated disciplines (incongruent) and racially/ethnically minoritized men in neutral disciplines (neutral). These findings indicate that while racially/ethnically minoritized role incongruent men are penalized by being viewed as harsh when in positions of authority, White men may be at an advantage and actually perceived more positively than other instructor groups. The opposite pattern is true for men teaching in neutral disciplines, racially/ethnically minoritized men are at an advantage due to their perceived

role neutrality while White men are penalized for not being on one side of the gender role spectrum or the other.

Two other themes emerged with respect to grades. Racially/ethnically minoritized men in man-dominated disciplines (congruent), racially/ethnically minoritized women in woman-dominated disciplines (congruent), and White women in man-dominated disciplines (incongruent) were described as not grading things in a timely manner. Additionally, White men in woman-dominated disciplines (incongruent), racially/ethnically minoritized men in neutral disciplines (neutral), and racially/ethnically minoritized women in neutral disciplines (neutral) were described as not putting grades on the courses' online learning management system. These findings are not clearly related to the instructors' gender, perceived gender role (in)congruity, or race/ethnicity but they do further illustrate that grading practices and grades are a key point of interest for students when evaluating their instructors.

Several instructor groups were described as "hard" or some variation of "hard". Racially/ethnically minoritized men man-dominated (congruent), racially/ethnically minoritized women woman-dominated (congruent), White men woman-dominated (incongruent), racially/ethnically minoritized men woman-dominated (incongruent), and racially/ethnically minoritized men neutral (neutral) were all described as "hard". Two other instructor groups were described as "hard to learn from" those being White women woman-dominated (congruent) and racially/ethnically minoritized men neutral (neutral). Additionally, four instructor groups were described as hard to understand those being racially/ethnically minoritized men man-dominated (congruent), racially/ethnically minoritized men woman-dominated (incongruent), racially/ethnically minoritized women man-dominated (incongruent), and racially/ethnically minoritized men neutral (neutral). Thus, racially/ethnically minoritized instructors were four of

the five groups described as “hard”, one of the two groups described as “hard to learn from”, and all four of the groups described as “hard to understand”. The four racially/ethnically minoritized instructor groups who are described as “hard to understand” may be described this way due to their accents or other language barriers as many of these instructors are not native English speakers. Due to sample size constraints in the available data, it was not possible to separate out racially/ethnically minoritized native and non-native English speakers. In order to tease out if racially/ethnically minoritized instructors are generally “hard to understand” or if it is due to language barriers, more research would need to be conducted in which racially/ethnically minoritized native English speakers and racially/ethnically minoritized non-native English speakers could be separated out.

The pattern of referring to racially/ethnically minoritized instructors as hard indicates that students may perceive the course requirements of their racially/ethnically minoritized instructors as being more challenging than those of their White instructors. Racially/ethnically minoritized women in neutral disciplines (neutral) were the only racially/ethnically minoritized group to not be referred to as some variation of “hard” while racially/ethnically minoritized men regardless of their level of congruity were referred to as at least one of three variations of “hard” listed. Racially/ethnically minoritized men in neutral disciplines (neutral) were the only neutral group to be referred to as any variation of “hard” and they were the group who was most frequently described using the word “hard” with students using phrases like: hard to know what they wanted, hard to learn from, hard to follow, and hard to understand. This is in stark contrast to racially/ethnically minoritized women in neutral disciplines (neutral), who were the only racially/ethnically minoritized group for whom the word “hard” was not used to describe them at all. This may be due to racially/ethnically minoritized women being viewed as less traditionally

feminine than White women but not sufficiently masculine and thus fitting better into the neutral discipline than their racially/ethnically minoritized man peers. Racially/ethnically minoritized men may be seen as overly masculine thus making it difficult for students to reconcile their masculine perception with the perceived gender ambiguous discipline in which they teach.

"Difficult" was also a word used to describe many instructor groups in a negative sense but for a couple of instructor groups the concept of "difficult" was used in a positive connotation. White men woman-dominated (incongruent), White women man-dominated (incongruent), racially/ethnically minoritized women man-dominated (incongruent), and White women neutral (neutral) were all described as "difficult". Furthermore, racially/ethnically minoritized men woman-dominated (incongruent), racially/ethnically minoritized women man-dominated (incongruent), and racially/ethnically minoritized men neutral (neutral) were described as having difficult materials. Students appear to be more likely to perceive gender role incongruent instructors as "difficult" with three of the four groups described this way but more likely to refer to racially/ethnically minoritized perceived gender role incongruent and neutral instructors as having "difficult materials" with all but racially/ethnically minoritized women described this way. On the other hand, racially/ethnically minoritized women in neutral disciplines (neutral) were described as being "not difficult" and White men in neutral disciplines (neutral) were described as having helped with difficult topics and being "difficult but worthwhile". The positive use of the word difficult to describe racially/ethnically minoritized women in neutral disciplines (neutral) is once again, much like the use of the word hard, in stark contrast to racially/ethnically minoritized men in neutral disciplines (neutral) and in this case also contrasts with White women in neutral disciplines (neutral). This combined finding further suggests that racially/ethnically minoritized women may not be held to the same gender role congruity

standards as their White women and racially/ethnically minoritized men peers. Racially/ethnically minoritized women especially in neutral disciplines may in fact be perceived and evaluated more similarly to White men in neutral disciplines who also were praised using the word "difficult".

On a gendered note, women instructors and racially/ethnically minoritized women instructors, in particular, were frequently described as being biased, not open to others' opinions, or condescending. Students perceived racially/ethnically minoritized women in woman-dominated disciplines (congruent) and neutral disciplines (neutral) to be biased and racially/ethnically minoritized women in woman-dominated disciplines (congruent) and White women in neutral disciplines (neutral) to not be open to others' opinions. However, white women in neutral disciplines (neutral) were also the only instructor group described as valuing students' opinions. Thus, it is racially/ethnically minoritized women in gender role congruent and neutral disciplines who are viewed as the most one-sided. This may be due to racially/ethnically minoritized women not being viewed as authority figures or not seen as being seen as subject matter experts even in disciplines in which women have historically dominated because that dominance has historically been by White women. Thus, racially/ethnically minoritized women's racial/ethnic identity may cause students to view them as less of an expert and not take their assertion of knowledge as seriously as other instructor groups.

White women in man-dominated disciplines (incongruent) were the only instructor group to be explicitly described as condescending. White women may not be viewed as a subject matter expert in man-dominated disciplines due to their perceived gender role incongruity with the subject. Students may not perceive them as positively when they assert their knowledge and therefore describe assertions of expertise as "condescending" because of the role mismatch.

Women and White women in particular are traditionally expected to be deferential and thus if a woman instructor did not defer to the knowledge of her students especially in a man-dominated discipline where not only is she perceived to lack-fit with the role but the students also tend to be primarily men, she may be likely to be perceived and evaluated as being condescending. Only one group of men instructors was described as "thinking they knew more than others" which were racially/ethnically minoritized men in woman-dominated (incongruent) disciplines. This reinforces the notion that students may not expect racially/ethnically minoritized persons to be subject matter experts and when they are, especially racially/ethnically minoritized men in a woman-dominated discipline, they are perceived negatively when they assert their knowledge and expertise.

Several instructor groups were also described as being "rude" by their students. White men man-dominated (congruent), racially/ethnically minoritized women woman-dominated (congruent), White man woman-dominated (incongruent), White women man-dominated (incongruent), and racially/ethnically minoritized men woman-dominated (incongruent) were all described as being "rude" by at least one student. White men, White women, and racially/ethnically minoritized men who deviate from their respective traditionally prescribed gender norms are described as "rude" while racially/ethnically minoritized women who are perceived to deviate from traditionally prescribed gender norms are not perceived this way. This indicates that while perceived gender role incongruence may lead to a higher likelihood of some instructor groups (White men, White women, and racially/ethnically minoritized men) being evaluated as "rude", this trend is gender and race/ethnicity dependent as racially/ethnically minoritized women are actually more likely to be evaluated as "rude" when they do in fact teach in perceived gender role congruent disciplines. Racially/ethnically minoritized women who teach

in woman-dominated disciplines may be more likely to be evaluated as "rude" because these disciplines while traditionally woman-dominated have historically been dominated by White women. Thus, while a racially/ethnically minoritized woman is perceived to be gender role congruent to the discipline, she is still violating the stereotypical image of a woman instructor in traditionally woman-dominated fields. Racially/ethnically minoritized women, therefore, may be punished due to this perceived lack of fit with the role of teaching in traditionally white woman-dominated disciplines. Racially/ethnically minoritized women may, however, be at an advantage in traditionally man-dominated disciplines because they are deviating from so many roles the gender role deviation is not considered as heavily by students when they are perceiving and evaluating the instructors. The only other perceived gender role congruent group who was described as "rude" is White men. White men in man-dominated disciplines may be more likely to not only feel like they are the expert and authority in the room but also to exert their expertise and authority which may result in perceptions and evaluations of rudeness by their students.

White perceived role incongruent men (woman-dominated) were the only group described as being a push over and "lacking care and respect for their students". White men in woman-dominated disciplines may be perceived as pushovers or not strong authority figures because of the feminine gendering of the discipline in which they teach. White men instructors teaching in a woman-dominated discipline may experiences some advantages relative to attaining promotions and leadership positions by riding what has been referred to as a "glass escalator" to these higher statuses (Williams 1992). However, in interpersonal situations men in traditionally feminine occupational roles may feel pressure to be extra masculine and authoritative in order to be viewed as appropriately masculine despite the femininity of their work domain (Simpson 2004b; Williams 1992). In the case of college instructors, teaching is a

very interpersonal and often considered feminine activity which may lead men instructors in role incongruent disciplines to feel pressure to act overtly masculine when interacting with student in order to overcompensate for the not only women-domination of the discipline in which they teach but also because they are teaching in general in order to ensure that their students view them as sufficiently masculine despite their general occupation and specific subject area expertise.

Furthermore, while at least one student referred to White men in women-dominated disciplines as "lacking care", at least one other student commented that their White man in a woman-dominated discipline did care. Some White men instructors in woman-dominated disciplines may be perceived as "lacking care" due to a higher expectation of instructors in woman-dominated disciplines to be caring and respectful because of the feminine characteristics associated with the discipline. Additionally, these disciplines are still heavily taught by White women who may exhibit more "caring" behaviors than men who teach in the discipline. Thus, when a man does not show enough caring, they may be called out for the lack of care required by the discipline with respect to both the historical and relative caring norms. When men do show enough caring, they are praised for meeting the historical and relative expectation of the discipline. White men in woman-dominated disciplines may be held to a higher caring standard than racially/ethnically minoritized men in woman-dominated disciplines because once again these woman-dominated disciplines were traditionally dominated by White women thus the expectation for caring remains stronger for men of the same racial/ethnic group than it does for others.

Caring and care more generally were major themes that arose across instructor groups in different ways. White men man-dominate (congruent), racially/ethnically minoritized women

woman-dominate (congruent), White men woman-dominate (incongruent), White women man-dominate (incongruent), racially/ethnically minoritized women man-dominate (incongruent), racially/ethnically minoritized men woman-dominate (incongruent), White men neutral (neutral), and White women neutral (neutral) were described as caring generally and/or caring for students. Racially/ethnically minoritized women in woman-dominated disciplines (congruent) were also noted as caring specifically about their students' learning and White women in neutral disciplines (neutral) were also described as "caring about the subject they teach". All but two perceived congruent groups were described as caring those being White women and racially/ethnically minoritized men.

White women in woman-dominated disciplines may not be described as caring because of the gendered expectation of White women to be caring and the traditionally White woman gendering of the discipline in which they teach. Thus, it may take an extraordinary amount of effort for White congruent women to be evaluated as caring because they are simply expected to be already. Racially/ethnically minoritized men may also have a difficult time being evaluated as caring due to racialized and gendered stereotypes of racially/ethnically minoritized men and especially Black men as being threatening or non-caring. Thus it may also take an extraordinary amount of effort for a racially/ethnically minoritized man to be evaluated as caring by their students which did not occur in the coded sample. Neither racially/ethnically minoritized men nor racially/ethnically minoritized women in neutral disciplines were described as caring. Students may already struggle to picture racially/ethnically minoritized persons in the role of college instructor and then when the gender ambiguity of the discipline is added on top of this, they may struggle even more to positively perceive the instructor. Thus, once again it may take

extraordinary effort on the part of racially/ethnically minoritized instructors in gender neutral disciplines to be perceived of and evaluated as caring by their students.

Only two instructor groups were called out for not caring, White men in woman-dominated disciplines (incongruent), as mentioned above, and racially/ethnically minoritized men in man-dominated disciplines (congruent). Once again, it may be that racially/ethnically minoritized men are not perceived of as fitting with the role of instructor even in man-dominated disciplines because these disciplines have been historically dominated by White men. Thus, when racially/ethnically minoritized men enter into the role of instructor even in a man-dominated discipline they are at a disadvantage as compared to their White man peers and must be extra caring and extra nice in order to receive the same level of compliments on student evaluations as their White men peers. Furthermore, as mentioned above while White men in feminine roles may experience a positive effect from their gender role incongruity known as the "glass escalator" (Williams 1992), it has been found that these same privileges are not extended to racially/ethnically minoritized men (Wingfield 2009). Racially/ethnically minoritized men and Black men in particular may not be perceived of as being "professional enough" to be in the role of professor let alone in a woman-dominated discipline (Wingfield 2009). Furthermore, racially/ethnically minoritized men and Asian men in particular may be perceived as lacking hegemonically masculine characteristics due to racial stereotypes and thus must very deliberately work to achieve the status of being viewed as "masculine" in most situations let alone in situations in which they enter into a traditionally feminine domain (Chen 1999). Thus, when White men teach in woman-dominated disciplines they may overcompensate for their perceived gender role incongruity by overemphasizing their masculine traits order to still be viewed as masculine leaders in their classroom and take advantage of the "glass escalator". However, for

racially/ethnically minoritized men overemphasizing masculine traits may not lead to the same increases in credibility as an authority figure as it does for White perceived role incongruent men. Racially/ethnically minoritized men overemphasizing their masculinity may, in the case of Black men, further alienate them from the feminine characteristics associated with the discipline thus leading students to view them as "uncaring" or, in the case of Asian men, overemphasizing their masculinity may not be enough to overcome the double stereotype of generally being perceived as less masculine than White men and being in a feminine domain.

White women in neutral disciplines were the only instructor group in which a student said their instructor "came across as bitchy". The use of the clearly gendered derogatory term "bitchy" to describe women in perceived gender neutral disciplines could be driven by the woman instructor not being seen as feminine enough while also not deviating far enough into a masculine authority role. These women in neutral disciplines may also be viewed as a threat to the gender hierarchy due to their existence in between the genders. Thus, White women in perceived neutral disciplines are perceived of by students as being in a middle neutral space where they are not viewed as sufficiently feminine or deviating sufficiently into masculine territories and thus they threaten the very foundation of binary gender roles and therefore any slightly negative tone or action could be perceived of negatively, or more specifically "bitchy".

Racially/ethnically minoritized instructors tended to be described as "unprofessional" when they were women teaching in woman-dominated disciplines (congruent), men teaching in woman-dominated disciplines (incongruent), and men teaching in neutral disciplines (neutral). This pattern is not clearly related to the instructors' levels of gender role (in)congruity, but it does suggest that students may not take racially/ethnically minoritized instructors seriously as professionals when they enter into the traditionally White role of college instructor. While this is

not the case for all racially/ethnically minoritized instructor groups, the groups who were not specifically referred to as "unprofessional" may also face these preconceived notions of not fitting with the role and therefore may need to go above and beyond or be especially credentialled in order to be viewed as professional by their students.

On the positive side, White men, White women, and racially/ethnically minoritized women were frequently described by their students as being "nice". Instructor groups described as "nice" included White men man-dominated (congruent), White women woman-dominated (congruent), racially/ethnically minoritized women woman-dominated (congruent), White men woman-dominated (incongruent), White women man-dominated (incongruent), racially/ethnically minoritized women man-dominated (incongruent), and racially/ethnically minoritized women neutral (neutral). White women were the only group in which a student described them as nice using gendered language by referring to their instructor as "a nice lady". The addition of the instructor's gender to the compliment of "nice" may be due to the student overemphasizing the feminine aspects of their perceived gender role incongruent woman instructor order to make sense of their perceived gender role incongruity through reemphasizing the fact that the instructor is a woman.

The only instructor group not described as "nice" at all were racially/ethnically minoritized men. White men and White women in neutral disciplines were also not described as being "nice". Students may not view White instructors in neutral disciplines as "nice" because of their gender ambiguity. While White incongruent instructors violate their gender role, they are still clearly expressing one gender role while White instructors in neutral disciplines sit in between traditionally prescribed gender roles which may be more difficult to cognitively process. For example, a White woman in a man-dominated discipline may be clearly violating gender

norms whereas a woman in a neutral discipline may be viewed as trying to toe the line and therefore be viewed even less positively than a gender role incongruent White woman. Racially/ethnically minoritized women in neutral disciplines may not have these same negative reactions to their perceived gender neutrality because they are already viewed as violating racial/ethnic norms by entering into traditionally White and traditionally man-dominated occupations and thus may be perceived as being "nice" even when they teach in gender role neutral disciplines. Racially/ethnically minoritized men may simply be viewed as entering into traditionally White occupations and therefore no matter the discipline they teach in, they are not perceived or evaluated to be "nice" by their students. This may especially be the case for Black men who are stereotypically portrayed and thought of as intimidating thus making it even more difficult for Black instructors to be perceived and evaluated as being "nice" by their students.

CONCLUSION

Overall, qualitative student evaluations much like quantitative student evaluations, tend to skew positive as supported by the code counts, sentiment analyses, and word clouds. Some quantitative measures of code counts do indicate that racially/ethnically minoritized women instructors tend to receive harsher evaluations than their White men, White women, and racially/ethnically minoritized men peers. This finding echoes that of previous SEI research which found that racially/ethnically minoritized instructors tend to receive harsher comments than White instructors (Aruguete et al. 2017). The quantitative code counts in this study add further nuance to the literature by highlighting not only the racial/ethnic disparity in student evaluation comments but also the gendered component of the comments.

The quantitative measures also indicate that institutions need to consider the goals of their student evaluations when selecting which open-response questions to include. The code

counts show that the questions on the old SEI forms (Course and Instructor) tend to evoke comments about the instructor themself while the questions on the new SEI forms (Helped Learn and Change) tend to evoke comments on the substance and materials of the course. Thus, if the goal of the institution is to solicit feedback specifically about the instructor, modeling questions on the Course and Instructor questions would be more useful. On the other hand, if the goal of the institution is to elicit feedback on the course itself, modeling questions on Helped Learn and Change would be the more useful approach. The choice of questions used should be driven by the goals of the institution and driven by the way in which the student evaluations are going to be used. These goals should be made explicitly clear to instructors when they are developing their courses to ensure that they are building classes with the institution's goals in mind and the goals should be made explicitly clear to students when they are completing their evaluations so that they can take a more informed approach to doing so.

Through further qualitative analyses, and especially comparison of the code summaries for each instructor group, some nuanced differences in the ways in which students write about their instructors emerged though on the whole students tend to talk about all instructor groups in similar ways. Women instructors in general were evaluated as harsh or unfair graders by their students. Women instructors may be viewed as being harsher graders because of students' perceptions of their role incongruity as an authority figure or women instructors may actually be harsher graders because they take more time and put more effort into grading than their men peers. The institution from which this data was collected is an R1 institution and therefore the faculty put a heavier emphasis on research than they would at other schools such as small liberal arts colleges. Due to this heavier emphasis on research and the traditionally feminized role of "teacher", women instructors may feel an obligation to and then actually put more emphasis on

their role as a teacher while men instructors put more emphasis on their role as a researcher. The disparity in effort put into teaching and especially grading may then lead to actual and perceived differences in grading by women versus men instructors. For example, if women instructors spend more time grading and therefore make more edits to and give more feedback on written work, even if that feedback is constructive, it may visually appear to be harsher due to the sheer amount of marking on the assignment. Thus, even when women instructors give good grades and good feedback, the simple act of marking up pages more than men instructors may make women instructors appear to be harder and therefore students evaluate them as such. Furthermore, students may have greater expectations of their women instructors being kinder and higher quality teachers because the role of teacher is traditionally feminine and thus students evaluate them more harshly when they do not meet their higher expectations while men instructors can do a lot less to receive just as good if not a better evaluation.

The difference in evaluation of women as unfair graders as well as other patterns of comments by instructor gender, perceived gender role (in)congruity, and race/ethnicity indicate that both RCT and SIH may both be useful theoretical approaches to examining how role congruity affects subjective evaluations. For example, on the one hand women instructors were more likely to evoke more visceral comments such as being called “bitchy” and “rude” more so than their men peers. However, men and especially racially/ethnically minoritized men also received negative comments especially when they were perceived to deviate from gender roles through being called “difficult” or “not caring”. Almost all perceived gender role incongruent instructors were referred to as “difficult” and racially/ethnically minoritized instructors were especially likely, in most, to receive negative comments with the word “difficult”.

The clearly mixed results of both women and men instructors receiving negative comments provides support for Status Incongruity Hypothesis (Moss-Racusin et al. 2010; Rudman et al. 2011). However, the lack of visceral responses and low disgust scores on the sentiment analyses indicate that while SIH is correct in that both men and women may be critiqued for perceived gender role incongruity or even neutrality, RCT may be more correct in the execution of the critique. Many instructors and especially women were noted as being difficult or hard which may have more to do with not seeing the women instructors as authorities or leaders of their classroom in general regardless of the gender dominance of the discipline (Eagly and Karau 2002). Scholars should therefore consider RCT's proposition that critiques of gender role congruity may be more likely to occur in leadership contexts but also SIH's proposition that backlash to perceived gender role incongruity may occur to people of all genders. The results of this study highlight that it is *both* of these causal mechanisms that affect subjective evaluations, it depends on what is being evaluated. Some questions are more likely to evoke moral outrage while some evoke more critiques of inadequacies in leadership or authority.

Limitations and Future Studies

In this study, 1,430 open-ended student evaluation responses were coded and themes were examined by instructor gender, perceived gender role congruity, and race/ethnicity. While this is a robust sample, not all unique identities could be accounted for due to the available data. For example, all racially/ethnically minoritized instructors were grouped into one racial/ethnic category due to low sample sizes of individual groups. While this provides some perspective on the differences in the ways in which students talk about White versus racially/ethnically minoritized instructors, more nuanced results would be achieved if more specific racial/ethnic categories could be examined. There is evidence that suggests that perceptions of gender role

congruity vary widely between different racial and ethnic groups (Biernat and Sesko 2013; Livingston et al. 2012) and while some differences were observed here, the causal mechanisms may be better determined with larger sample sizes for more specific racial and ethnic categories such as Latinx, Asian, etc. For example, in this study it was found that racially/ethnically minoritized instructors were more likely to be described as some version of "hard" which may be due to language barriers which may be more common among racially/ethnically minoritized instructors due to English being their second language. However, the data did not indicate which instructors spoke English as their second language and which instructors spoke English as their first language thus this nuance could not be systematically parsed out. Future studies should attempt to solicit greater sample sizes for more nuanced racial and ethnic categories as well as other potentially salient identities such as speaking English as second language, LGBTQ+ status, tenure at the institution, and age. These salient identities and more may affect students' perceptions and evaluations of their instructors.

Furthermore, non-response rates may skew the comments instructors receive on their student evaluations of instruction. The comments coded in this study tended to be more positive than negative and the quantitative analyses from the previous two studies (Chapters 2 and 3) further indicate the positive skew of the evaluations. The positive skew of the data may be due to many students with average or below average perceptions of the instructors simply not completing the student evaluation form. It is an opt-in process that is not required and completion is not consistently incentivized therefore there may be little motivation for completion. Future studies should examine the characteristics of non-responders versus responders to determine if there are any fundament differences in the populations which may be leading to the observed skew in the results.

Finally, while many of the comments may have been similar across faculty groups, the comments may have differential effects on instructors and how they think about their role as an instructor. It would be useful to interview, observe instructors, or have instructors complete daily time-use journals to determine if there are gender, racial/ethnic, and/or perceived gender role (in)congruity differences in the ways in which instructors interpret, internalize, and act upon student evaluations generally and especially qualitative open-ended comments. Interviews, observations of faculty, or time-use journals would also be useful to determine if, as previous research suggests (e.g. El-Alayli et al. 2018), faculty of different genders, races/ethnicities, and/or levels of role (in)congruity are putting forth different levels of effort in order to receive similar qualitative feedback. Students may not realize that some faculty are going above and beyond to be perceived of as professional instructors but time-use journals, observations, and interviews of instructors may be able to determine if this is the driving force behind different instructor groups receiving similar qualitative evaluations. Interviews or observations of instructors may also determine the merit of comments in which different groups of instructors do appear to be perceived and evaluated differently such as how "hard" or "difficult" they are as compared to their peers. If it is found that instructors are not harder but rather it is students' perceptions of difficulty which are biased, student evaluation of instruction forms could be further refined in order to mitigate as much differential treatment as possible.

Closing Remarks

As found in this study, qualitative student evaluation items may not differ greatly between different groups of instructors based on their gender, gender role (in)congruity, and/or race/ethnicity but they vary enough that it is important to be reflexive and refine the measures so that they are aligned with the goals of the institution and are as equitable as possible to faculty of

all identities. Institutions need to be clear in their evaluation goals and provide support for instructor groups who may be unfairly evaluated or unfairly worked in order to receive the same evaluations in order to mitigate these discrepancies as much as possible. Furthermore, the results of this study indicate that future studies should consider the propositions of both RCT and SIH in tandem when examining the effect of perceptions of gender role (in)congruity on subjective evaluations as well as consider other statuses which may affect perceptions such as the race/ethnicity of the target of the evaluation.

CHAPTER 5: SUMMARY AND CONCLUSION

SUMMARY OF RESEARCH

In this dissertation, three research studies were completed to examine a potential source of bias in the content and completion of subjective evaluations—reactions to perceived target role (in)congruity. Specifically, I examined whether instructors' gender, race/ethnicity, and the gender role dominance of the discipline in which they teach affects how students evaluate them on student evaluations of instruction (SEIs). I sought to answer the following questions: (1) Are SEIs affected by faculty perceptions of gender role (in)congruity? (2) Does the effect of perceptions of faculty gender role congruity also depend on the race/ethnicity of the faculty? To answer these questions, multi-method analyses were completed including quantitative and qualitative analyses of traditional close- and open-ended student evaluation measures.

In these studies, theories of congruity, specifically Role Congruity Theory (RCT) and Status Incongruity Hypothesis (SIH) were also examined, and their current theoretical predictions were applied in new contexts. To date, theories of congruity have focused exclusively on gender and have not considered how other salient social roles may affect perceptions of gender role (in)congruity. Additionally, theories of congruity have not examined how persons in perceived gender neutral roles are affected by perceptions of target role congruity. In this dissertation, the effect of perceived gender role neutrality and the interaction of race/ethnicity with gender and perceived gender role (in)congruity were examined. Recommendations for additions and changes to current theories of congruity can be made based on the results.

Finally from a practical perspective, the current quantitative student evaluation questions of an institution of higher education were examined to determine the most appropriate

measurement model. Furthermore, two sets of qualitative student evaluation questions were tested to determine if the different questions led to different evaluations. Based on the results of the three studies, recommendations for optimizing student evaluations to collect data that best achieves the goals of the institution can be made.

Study 1 (Chapter 2): Quantitative Analyses of Student Evaluations of Instruction with Attention to Faculty Gender and Gender Role (In)Congruity

In Study 1 (Chapter 2), student evaluations of instruction were quantitatively analyzed with consideration of the gender and perceived gender role (in)congruity of the course instructor. Through the analyses, I sought to answer the question are students' subjective evaluations of their instructors affected by the perceived gender role (in)congruity of the instructor? Confirmatory factor analyses, MIMIC models, and grouped structural equation models were used in the analyses. Results of the analyses indicate that it is crucial to test the appropriateness of a measurement model prior to testing for differences in SEI scores between men and women instructors. For all three groups of faculty - role congruent, role incongruent, and role neutral - variables were determined to be measured differently depending on the gender of the instructor. This result means that simply comparing the means or regression tests on the unconstrained model would have been biased due to the measurement errors of the model itself and the results would be unreliable. In Study 1 (Chapter 2), the proper constraints were added to the model for each of the three instructor groups so that accurate comparisons of the mean scores of men and women in each group could be compared.

Once the proper constraints were added to the models for perceived gender role incongruent and perceived gender role neutral faculty, there were no longer statistically significant differences in the means of either latent SEI construct, Overall or Instructor. For perceived role congruent faculty, statistically significant differences in the means of the latent

constructs persisted even when the model was properly constrained. Thus, student evaluations of instruction may be more sensitive to differences in students' perceptions of their instructors when instructors teach in perceived role congruent disciplines than when they teach in perceived role incongruent or role neutral disciplines. Students may perceive instructors in gender role congruent disciplines to be experts in the field due to their perceived gender role congruity and therefore be more critical of the quality of the course and instructor. Students may not have as high of expectations for perceived role incongruent or perceived role neutral faculty because of the perceived mismatch of their gender with the discipline in which they teach. Further research should work to determine the causal mechanism behind the continued measurement invariance for perceived role congruent men and women instructors and the lack of continued measurement invariance for perceived role incongruent and perceived role neutral men and women instructors.

Study 2 (Chapter 3): Quantitative Analyses of Student Evaluations of Instruction with Attention to Faculty Gender, Race/Ethnicity, and Gender Role (In)Congruity

In Study 2 (Chapter 3), student evaluations of instruction were quantitatively analyzed with consideration of the gender, race/ethnicity, and perceived gender role (in)congruity of the course instructor. Through these analyses, I sought to build on the results of Study 1 to answer the question are students' evaluations not only affected by perceived instructor role (in)congruity *but also* the race/ethnicity of the instructor? Through the addition of race/ethnicity, this study pushes previous research using theories of congruity by determining if another salient social characteristic, in this case race/ethnicity, further affects perceptions and evaluations of persons in addition to their perceived level of gender role (in)congruity. Once again, grouped structural equation models were used in the analyses. The results of Study 2 (Chapter 3) further highlight the importance of testing the appropriateness of a measurement model prior to testing for

differences in the mean student evaluation scores for different instructor groups as much of the differences between groups were eliminated once models were weighted appropriately.

The results of Study 2 (Chapter 3) do indicate that when measurement invariance is accounted for and instructor race/ethnicity is considered in tandem with instructor gender and perceived level of gender role congruity, some of the between-group differences in SEI scores persist depending on the role congruity group being examined. For perceived gender role congruent faculty, statistically significant differences on the latent concept Overall persisted even when measurement invariance was accounted for such that the scores for White women were the highest and the scores for racially/ethnically minoritized women were the lowest. For perceived gender role incongruent faculty, once the models were constrained based on the determined measurement invariance there were not statistically significant differences in the means of either latent construct, Overall and Instructor, regardless of the race/ethnicity of the instructor. Finally, for perceived gender role neutral faculty, statistically significant differences on the latent concept Instructor persisted even when measurement invariance is accounted for such that White men scored higher than all other groups and racially/ethnically minoritized men scored lower than all other groups. Thus, the results of this study and Study 1 (Chapter 2) highlight not the importance of testing and accounting for measurement invariance and the results of this study further highlight the need to consider persistent differences in measurement invariance that remain when level of perceived gender role (in)congruity and another salient social characteristic, in this case race/ethnicity, are considered even when proper constraints are applied to the measurement models.

Study 3 (Chapter 4): Qualitative Analyses of Student Evaluations of Instruction with Attention to Faculty Race/Ethnicity, Gender, and Gender Role (In)Congruity

In Study 3 (Chapter 4), the results of Studies 1 and 2 (Chapters 2 and 3) were further built upon as student evaluations of instruction were qualitatively analyzed with consideration of the gender, perceived gender role (in)congruity, and race/ethnicity of the course instructor. Through these analyses, I sought to answer the questions: are students' *open-ended* subjective evaluations of their instructors affected by the race/ethnicity and/or perceived gender role (in)congruity of the instructor; what is the motivation behind potential differences in subjective evaluations of instructors based on their perceived level gender role (in)congruity, and do the types of qualitative student evaluation questions asked have different effects on students' subjective evaluations of their instructors? Open-response student evaluation questions were coded across six code categories which were then quantitatively and qualitatively analyzed.

Both the quantitative and qualitative results from code counts, sentiment analyses, word clouds, and thematic processing were overwhelmingly positive. Some of the quantitative code counts indicate that there are differences in how and how much students talk about their instructors which depends on the gender and/or race/ethnicity of the instructor. Racially/ethnically minoritized women were more likely to receive harsher evaluations than their racially/ethnically minoritized men, White men, and White women peers. Additionally, all women instructors were more likely than men instructors to be evaluated as "harsh" and "unfair graders". However, men and especially racially/ethnically minoritized men received negative comments when they deviated from gender roles such as being called "difficult" or "not caring". Almost all perceived gender role incongruent instructors were referred to as "difficult" and racially/ethnically minoritized instructors were especially likely, in most, to receive negative comments with the word "difficult".

These results add nuance to the findings of previous research by highlighting that there are gender, racial/ethnic, perceived gender role, and combined gender-racial/ethnic-perceived gender role differences in the ways in which students qualitatively evaluate their instructors. Furthermore, the quantitative results in particular show that the type of question asked by an institution can highly affect the types of comments students write. When questions were very broad and simply asked for "comments on instructor" or "comments on course", students tended to write more personal and professional comments about the instructor while when SEIs asked more specific questions about "what helped you learn" or "what recommendations do you have for change", students tended to write comments about the course and its content rather than the instructor. Institutions need to be very min*df*ul of the goals of their student evaluations of instruction and select open-response questions that will be most likely to get responses from students which meet these goals.

CONCLUSION

Taken together, the results of these three studies show the importance of considering perceptions of target role congruity when examining the results of subjective evaluations. The results of these studies have many theoretical and practical implications. While there are limitations to these studies, they also open up new avenues of research for both theories of congruity and student evaluations of instruction.

Theoretical Implications

From a theoretical perspective, the results of the three studies indicate that Role Congruity Theory and Status Incongruity Hypothesis are both useful theories but that there are ways in which they could be expanded in future research. The results of these studies indicate that perceptions of gender role congruity can lead to backlash for both men and women who are

role incongruent as proposed by SIH but that these backlash reactions may not necessarily come out as expressions of disgust. Furthermore, backlash reactions may be driven by moral outrage due to defiance of the gender hierarchy as predicted by SIH but more research is needed to fully support this claim.

While testing the current claims of RCT and SIH, the three studies presented here also add to previous research on how perceptions of role (in)congruity affect subjective evaluations. Firstly, these three studies indicate that other salient social roles may affect subjective evaluations in addition to gender role congruity. According to the results of Studies 2 and 3 (Chapters 3 and 4, respectively), the race/ethnicity of the target can affect how others subjectively evaluated them. The quantitative results found that perceived gender role congruent White instructors were rated more positively on the latent concept Overall than racially/ethnically minoritized instructors with racially/ethnically minoritized women receiving the lowest scores. The race/ethnicity of instructors did not affect the scores of perceived gender role incongruent instructors, but for gender role neutral faculty racially/ethnically minoritized instructors and racially/ethnically minoritized men were rated especially lower than their White peers. The qualitative results indicate that racially/ethnically minoritized instructors are more likely to receive certain types of negative feedback such as being referred to as variations of "hard" more often than their White peers. While the qualitative results may be due, in part, to language barriers some racially/ethnically minoritized instructors may have with their students, taken together the quantitative and qualitative results indicate that race/ethnicity can work in tandem with gender and perceived level of gender role congruity to affect subjective evaluations. This finding indicates that future work using theories of congruity should take into consideration how other salient social roles such as race/ethnicity may affect perceptions and evaluations.

Furthermore, it may even be useful for future research to consider if there are other salient social characteristics such as race/ethnicity that should affect what is even considered "congruent" in the first place. By incorporating more salient social characteristics, future studies of congruity may be more nuanced and accurately portray what occurs in real-life situations on a daily basis.

Secondly, the effect of perceived gender neutral roles were tested to determine if targets in such positions are affected by perceptions of role congruity. The results indicate that instructors in perceived gender neutral disciplines and especially racially/ethnically minoritized instructors were particularly vulnerable to especially negative qualitative evaluations. The majority of research using RCT and SIH have not, to my knowledge, previously examined how perceptions of gender neutral persons affect subjective evaluations, with one notable exception (Cabrera, Sauer, and Thomas-Hunt 2009). The results of this dissertation and especially that of Study 3 indicate that perceived gender role neutral positions can also cause backlash reactions which are, in some cases, even more extreme than backlash reactions to persons who occupy perceived gender role incongruent positions. Perceived gender role neutral instructors were the only groups not described as "nice" which may be due to the ambiguity of the lack of clear gender dominance of the discipline in which they teach. Though gender and perceived gender roles today are generally acknowledged to be spectrums with various levels of femininity, masculinity, and androgyny, many people still process gender in a very black-and-white, masculine-and-feminine manner. Due to the traditionally binary nature of gender and gender roles, people may struggle to process people who occupy roles that sit in between the two extremes which may lead to even more negative evaluations than even role incongruent persons receive. The results from Cabrera et al. (2009) further support these results as they found that in gender neutral contexts female leaders were rated statistically significantly higher than male

leaders. However, their results further indicate that there is not a statistically significant difference in the rating of female and male leaders on team performance in perceived gender neutral contexts (Cabrera et al. 2009). The results of Study 3 (Chapter 4) and other research (Cabrera et al. 2009) indicate that more research using theories of congruity is necessary to examine the nuanced effects of perceived gender neutral roles on perceptions and evaluations in order to determine if the results found in this dissertation are specific to the context of student evaluations or more widespread.

Finally, these three studies go a step beyond previous studies using theories of role congruity by examining the effects on both quantitative and qualitative evaluations in the same context. While there are previous studies in congruity of both quantitative and qualitative data there is not, to my knowledge, previous work which has examined both data types in one context. By examining the effects of perceptions of target role congruity on both quantitative and qualitative subjective evaluations in one context, these studies provide not only robust statistical evidence to support the results but also qualitative results which work to determine the causal mechanisms behind the quantitative analyses. Thus this dissertation constitutes a more robust examination of the causes and consequences of perceptions of target role (in)congruity on quantitative and qualitative subjective evaluations than has been previously conducted.

Given the findings of these three studies, it may be useful to not only extend the scope of RCT and SIH but to perhaps propose a new theory altogether that accounts for the propositions of both of these theories as well as the additions described here including examining the effect of other salient social roles, including gender neutral roles in analyses, and using both quantitative and qualitative research methods. A theory which considers how perceptions of different levels of gender role congruity, incongruity, and neutrality affect persons of all gender and how these

roles interact with other salient social roles would provide a more nuanced approach to studying the effects of perceived gender role (in)congruity on subjective evaluations than current theories of congruity. Furthermore, utilizing multiple research methods in a variety of contexts will help to determine if the effects of perceptions of role (in)congruity are contextually dependent. A new all-encompassing theory that takes the findings of this dissertation into account would provide a stronger foundation for future studies of the effects of role congruity on subjective evaluations.

Practical Implications

The results of the three studies presented in this dissertation highlight that the results of student evaluations of instruction can vary by the gender, perceived gender role congruity, and race/ethnicity of the instructor being evaluated but, in general, student evaluation scores are positive for both quantitative and qualitative measures. Furthermore, when measurement invariance is accounted for, differences between instructor groups are, for the most part, mitigated. However, in most cases, there was measurement invariance between different instructor groups. Therefore, institutions of higher education need to complete measurement invariance testing and properly weight student evaluation items before comparing the evaluation scores of different instructor groups. If institutions do not complete this step, they will be comparing apples to oranges and therefore not fairly controlling for between-group differences in scores. By completing measurement invariance testing and adding the appropriate weights to the model, institutions will be able to compare one type of apple to another thus providing much more accurate and appropriate between-group comparisons.

The results of these studies also indicate that different student evaluation questions produce very different results, especially with respect to qualitative open-response questions as indicated by Study 3 (Chapter 4). Institutions need to thoroughly consider the goal and purpose

of student evaluations when selecting which questions to include and then select questions to achieve their desired goals. The institution from which data was analyzed changed their questions during the time period analyzed. In the older time period the institution asked blanket questions that simply left open space for "comments on instructor" or "comments on course" which tended to lead to more comments about the personality or professionalism of the instructor themselves. When the institution changes to more directed questions that asked for "recommendations for change" or "what helped you learn" they tended to receive more comments about the course and its content. Thus, if an institution's goals for student evaluations are to solicit feedback about the instructor those goals would be better met with vague open-response prompts while if the institution's goals for student evaluations are to get feedback about the course, more targeted questions better achieve this goal. If institutions want to get feedback about the instructor and the course content, using a combination of specific and broad questions would best achieve this goal.

Quantitative questions also need to be carefully considered both separately and how they combine to measure different latent concepts. Institutions should use factor analyses to determine how their current observed variables combine in order to determine if all current observed variables are needed and/or if there are new observed variables that should be added to future evaluations. Therefore it is important for institutions to consider their evaluation goals and test the effects of their current measures in order to determine if any changes need to be made in order to better achieve their research goals.

Furthermore, institutions need to make both instructors and students aware of the purpose and goal of student evaluations. With respect to instructors, making the institutional goals of student evaluations clear may affect how they structure their courses and their teaching style. If it

is made clear that students will be explicitly asked about the content they found most useful, instructors may be more thoughtful about the content as they prepare to teach the course. Additionally, if it is the goal of the institution for instructors to make actionable changes to improve their teaching based on student evaluations, metrics should be put into place to measure any changes made in response to student evaluation feedback. One way to achieve this is for instructors to write about the changes they made as a result of SEIs in their annual evaluations including the specific steps they took to address any weaknesses in their teaching as identified in their evaluation results. Once again, if this is an expectation of the institution it needs to be made explicitly clear to instructors with sufficient time for them to make any identified changes necessary to improve their courses.

With respect to students, people today, especially high technology users such as younger traditionally college-aged persons, are bombarded with satisfaction surveys and the ability to review everything from the places they shop to the restaurants they eat in to professional drivers on the road with "How's my driving?" stickers on their bumpers. Due to the oversaturation of platforms to provide subjective evaluations, students may be unclear as to the purpose and goal of student evaluations. If the goal of student evaluations is to provide instructors with feedback to improve their teaching generally and the aspects of the specific course they took, this goal needs to not only influence the questions that are asked but also made explicitly clear to students. With more specific direction as to why they are completing student evaluations of instruction, students may provide more useful feedback. Additionally, if students are made aware of potential subconscious biases that may influence their evaluations, they may put forth effort to avoid allowing these biases to control their evaluations. Because gender, gender role (in)congruity, and racial/ethnic biases are likely not explicit choices students are making, raising awareness of the

effects of these biases could help to mitigate the between group differences observed in this dissertation.

Limitations

While this study has many theoretical and practical implications, it is not without its own limitations. Firstly, data was only from one institution in which the student body is overwhelmingly White and majority male, an anomaly in modern higher education. Given that instructor gender and race/ethnicity were two of the core variables examined here, there may be reason to believe that the results may be different at a more racially/ethnically diverse institution or at an institution with a majority female student body. According to previous research, people who are role incongruent themselves tend to rate others who are also role incongruent more positively than role congruent evaluators (Diekman and Schneider 2010b). Thus, there is reason to believe that racially/ethnically minoritized students and/or female students may rate their racially/ethnically minoritized and/or female instructors more positively than White male students. It would be useful if future research was conducted in a different institutional context and if future research was able to take into account the gender, race/ethnicity, and major of the student evaluators.

Similarly, the instructor population at the institution from which the data were obtained is also majority White. Due to the overwhelmingly White instructor population especially in role neutral disciplines, it was difficult to select courses that were of similar levels and in the same discipline across all instructor groups. Thus, completing the same analyses at a different institution in which the instructor population is more diverse would be useful to help determine if course level or specific disciplines create more nuanced differences in student evaluation scores.

Furthermore, all racially/ethnically minoritized instructors needed to be combined in order to have sufficient sample sizes across discipline type and instructor gender. The results would be more robust and descriptive if instructors of different race/ethnicities did not need to be combined into one category. Student evaluations of Black instructors may be different from evaluations of Asian instructors which may be different from evaluations of Latinx instructors and so on and within all of these groups there may be variations by gender and level of gender role (in)congruity. Testing between group differences of student evaluations of instructors at an institution with a more diverse instructor population may produce widely different results than those of the studies in this dissertation.

Additionally, it would be useful if future studies could compare the scores of lower- and higher-level courses. The data used in these studies did not have sufficient sample sizes of the different instructor groups to conduct these more nuanced results. However, there is reason to believe there may be significant differences in the scores of higher- and lower-level courses. These differences may be due to students in higher-level courses being more likely to be majors in the discipline and therefore they may be more invested in the course and therefore more critical in their evaluations than students in lower-level courses who may be taking a course simply to fulfill a university requirement and thus do not particularly care about the course or its quality.

Finally, both the quantitative and qualitative student evaluations used in these studies were overwhelmingly positive. Not all subjective evaluations may follow this trend. It would be useful if future research were to study the effect of target gender, gender role congruity, and/or race/ethnicity on subjective evaluations in other contexts which may have more diverse evaluations. A wider spread of evaluation results may lead to more between group differences

and stronger effects. One context which may provide more varied results is subjective evaluations of politicians as politicians are generally quite polarizing thus leading to more polarized results. Examining the effects of politicians' gender, level of gender role (in)congruity, and/or race/ethnicity on subjective evaluations may produce very different results and add to our understanding of the effects of role congruity in more polarizing contexts.

Future Studies

In addition to the future research mentioned in the limitations section, there are a multitude of other future research projects that could be explored to add to both research on role congruity and student evaluations of instruction. As mentioned above, future studies in role congruity should be conducted which consider the effects of other salient social roles, the effects of gender neutral roles, the effects of perceptions of role congruity on both quantitative and qualitative subjective evaluations, and the effects of perceptions of role congruity in more polarizing contexts. More research should be done to determine the causal mechanisms behind backlash reactions and to determine if different contexts result in different types of backlash. Experimental research to isolate the causal mechanisms may be useful in addition to qualitative research to delve into evaluators' thought-processes when they engage in backlash to role incongruent targets.

As mentioned above with respect to research on student evaluations of instruction, research should be done at institutions with more student and faculty diversity and completed to determine the effects of students' level of role (in)congruity on the ways in which they perceive and evaluate their instructors. Other future student evaluation research could include interviews with instructors to determine how they may or may not internalize the results of student evaluations and how student evaluations affect their teaching. Instructor-centered research may

also include observations of teaching and/or having instructors complete time journals to compare work loads across different instructor groups. Future student evaluation research may also include testing different quantitative and qualitative student evaluation measures to better optimize current student evaluation forms to better meet the assessment goals of the institution. Additionally, student-centered research such as interviews of how students think about student evaluations when they are completing them and to ask more detailed questions about their specific courses and instructors may be useful.

Concluding Remarks

According to the results of the three studies completed in this dissertation, perceptions of target gender role (in)congruity can affect both quantitative and qualitative subjective evaluations and these effects can be impacted by targets' other salient social roles besides gender. While these studies have their own limitations, they add significantly to previous research on perceived role congruity and student evaluations of instruction. Student evaluations are almost ubiquitous in higher education and subjective evaluations are ever more present in modern society. Thus, it is crucially important to understand how perceptions of the targets of these evaluations may affect the evaluation results. More research is needed to continue to develop theories of congruity and to create stronger less-biased student evaluations of instruction, but this dissertation constitutes major strides forward in both of these research areas.

BIBLIOGRAPHY

Acock, Alan C. 2013. *Discovering Structural Equation Modeling Using Stata*. StataCorp LP.

Algozzine, Bob, John Gretes, Claudia Flowers, Lisa Howley, John Beattie, Fred Spooner, Ganesh Mohanty, and Marty Bray. 2004. "Student Evaluation Of College Teaching: A Practice In Search Of Principles." *College Teaching* 52(4):134–41. doi: 10.3200/CTCH.52.4.134-141.

Anderson, Kristin J., and Gabriel Smith. 2005. "Students' Preconceptions of Professors: Benefits and Barriers According to Ethnicity and Gender." *Hispanic Journal of* B*ehavioral Sciences* 27(2):184–201. doi: 10.1177/0739986304273707.

Andreoletti, Carrie, Jennifer P. Leszczynski, and William B. Disch. 2015. "Gender, Race, and Age: The Content of Compound Stereotypes Across the Life Span." *The International Journal of Aging and Human Development* 81(1–2):27–53. doi: 10.1177/0091415015616395.

Anon. 2019. "Reconsidering Student Evaluations of Teaching." *American Sociological Association*. Retrieved June 22, 2020 (https://www.asanet.org/press-center/press-releases/reconsidering-student-evaluations-teaching).

Anon. 2021. "STATA Structural Equation Modeling Reference Manual Release 17."

Arbuckle, Julianne, and Benne D. Williams. 2003. "Students' Perceptions of Expressiveness: Age and Gender Effects on Teacher Evaluations." *Sex Roles; New York* 49(9/10):507–16. doi: http://dx.doi.org.www.libproxy.wvu.edu/10.1023/A:1025832707002.

Aruguete, Mara S., Joshua Slater, and Sekela R. Mwaikinda. 2017. "The Effects of Professors' Race and Clothing Style on Student Evaluations." *Journal of Negro Education* 86(4):494–502.

Basow, Susan A. 1995. "Student Evaluations of College Professors: When Gender Matters." *Journal of Educational Psychology* 87(4):656–65. doi: 10.1037/0022-0663.87.4.656.

Basow, Susan A., and Suzanne Montgomery. 2005. "Student Ratings and Professor Self-Ratings of College Teaching: Effects of Gender and Divisional Affiliation." *Journal of Personnel Evaluation in Education* 18(2):91–106. doi: 10.1007/s11092-006-9001-8.

Basow, Susan A., Julie E. Phelan, and Laura Capotosto. 2006. "Gender Patterns in College Students' Choices of Their Best and Worst Professors." *Psychology of Women Quarterly* 30(1):25–35. doi: 10.1111/j.1471-6402.2006.00259.x.

Bavishi, Anish, Juan M. Madera, and Michelle R. Hebl. 2010. "The Effect of Professor Ethnicity and Gender on Student Evaluations: Judged before Met." *Journal of Diversity in Higher Education* 3(4):245–56. doi: 10.1037/a0020763.

Benton, Stephen L., and William E. Cashin. 2014. "Student Ratings of Instruction in College and University Courses." Pp. 279–326 in *Higher Education: Handbook of Theory and Research*. Vol. 29, *Higher Education: Handbook of Theory and Research*, edited by M. B. Paulsen. Dordrecht: Springer Netherlands.

Biernat, Monica, and Amanda K. Sesko. 2013. "Evaluating the Contributions of Members of Mixed-Sex Work Teams: Race and Gender Matter." *Journal of Experimental Social Psychology* 49(3):471–76. doi: 10.1016/j.jesp.2013.01.008.

Blackburn, Heidi. 2017. "The Status of Women in STEM in Higher Education: A Review of the Literature 2007–2017." *Science & Technology Libraries* 36(3):235–73. doi: 10.1080/0194262X.2017.1371658.

Boring, Anne, Kellie Ottoboni, and Philip Stark. 2016. “Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness.” *ScienceOpen Research*. doi: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1.

Brescoll, Victoria L., Tyler G. Okimoto, and Andrea C. Vial. 2018. “You’ve Come a Long Way…Maybe: How Moral Emotions Trigger Backlash Against Women Leaders: Moral Emotions in Backlash.” *Journal of Social Issues* 74(1):144–64. doi: 10.1111/josi.12261.

Burn, Shawn Meghan. 1996. “Gender as a Social Category: Gender Differences and Errors in Thinking.” Pp. 109–35 in *The Social Psychology of Gender*. New York, NY: McGraw-Hill.

Cabrera, Susan F., Stephen J. Sauer, and Melissa C. Thomas-Hunt. 2009. “The Evolving Manager Stereotype: The Effects of Industry Gender Typing on Performance Expectations for Leaders and Their Teams.” *Psychology of Women Quarterly* 33(4):419–28. doi: 10.1111/j.1471-6402.2009.01519.x.

Cao, Chunhua, Eun Sook Kim, Yi-Hsin Chen, John Ferron, and Stephen Stark. 2019. “Exploring the Test of Covariate Moderation Effects in Multilevel MIMIC Models.” *Educational and Psychological Measurement* 79(3):512–44. doi: 10.1177/0013164418793490.

Centra, John A., and Noreen B. Gaubatz. 2000. “Is There Gender Bias in Student Evaluations of Teaching?” *The Journal of Higher Education* 71(1):17–33. doi: 10.2307/2649280.

Chávez, Kerry, and Kristina M. W. Mitchell. 2020. “Exploring Bias in Student Evaluations: Gender, Race, and Ethnicity.” *PS: Political Science & Politics* 53(2):270–74. doi: 10.1017/S1049096519001744.

Chen, Anthony S. 1999. “Lives at the Center of the Periphery: Chinese American Masculinities and Bargaining with Hegemony.” *Gender & Society* 13(5):584–607. doi: 10.1177/089124399013005002.

Clayson, Dennis E. 2009. “Student Evaluations of Teaching: Are They Related to What Students Learn?: A Meta-Analysis and Review of the Literature.” *Journal of Marketing Education* 31(1):16–30. doi: 10.1177/0273475308324086.

Davis, Leslie, and Richard Fry. 2019. “College Faculty Have Become More Racially and Ethnically Diverse, but Remain Far Less so than Students.” *Fact Tank: News in the Numbers*, July 31.

Diekman, Amanda B., and Alice H. Eagly. 2008. “Of Women, Men and Motivation: A Role Congruity Account.” Pp. 434–47 in *Handbook of motivation science*.

Diekman, Amanda B., and Monica C. Schneider. 2010a. “A Social Role Theory Perspective on Gender Gaps in Political Attitudes.” *Psychology of Women Quarterly* 34(4):486–97. doi: 10.1111/j.1471-6402.2010.01598.x.

Diekman, Amanda B., and Monica C. Schneider. 2010b. “A Social Role Theory Perspective on Gender Gaps in Political Attitudes.” *Psychology of Women Quarterly* 34(4):486–97. doi: 10.1111/j.1471-6402.2010.01598.x.

Diemer, Matthew A., and Cheng-Hsien Li. 2011. “Critical Consciousness Development and Political Participation Among Marginalized Youth: Critical Consciousness and Political Engagement.” *Child Development* 82(6):1815–33. doi: 10.1111/j.1467-8624.2011.01650.x.

Diemer, Matthew A., Adam M. Voight, Aixa D. Marchand, and Josefina Bañales. 2019. “Political Identification, Political Ideology, and Critical Social Analysis of Inequality among Marginalized Youth.” *Developmental Psychology* 55(3):538–49. doi: 10.1037/dev0000559.

Eagly, Alice H., and Steven J. Karau. 2002. “Role Congruity Theory of Prejudice toward Female Leaders.” *Psychological Review* 109(3):573–98. doi: 10.1037/0033-295X.109.3.573.

Eagly, Alice H., Wendy Wood, and Amanda B. Diekman. 2000. “Social Role Theory of Sex Differences and Similarities: A Current Appraisal.” Pp. 123–74 in *The Developmental Social Psychology of Gender*, edited by T. Eckes and H. M. Trautner. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

El-Alayli, Amani, Ashley A. Hansen-Brown, and Michelle Ceynar. 2018. “Dancing Backwards in High Heels: Female Professors Experience More Work Demands and Special Favor Requests, Particularly from Academically Entitled Students.” *Sex Roles* 79(3–4):136–50. doi: 10.1007/s11199-017-0872-6.

Falkoff, Michelle. 2018. “Why We Must Stop Relying on Student Ratings of Teaching.” *The Chronicle of Higher Education*, April 25.

Fassiotto, Magali, Jie Li, Yvonne Maldonado, and Nishita Kothary. 2018. “Female Surgeons as Counter Stereotype: The Impact of Gender Perceptions on Trainee Evaluations of Physician Faculty.” *Journal of Surgical Education* 75(5):1140–48. doi: 10.1016/j.jsurg.2018.01.011.

Foschi, Martha, Larissa Lai, and Kirsten Sigerson. 1994. “Gender and Double Standards in the Assessment of Job Applicants.” *Social Psychology Quarterly* 57(4):326. doi: 10.2307/2787159.

Fox, Richard L., and Zoe M. Oxley. 2003. “Gender Stereotyping in State Executive Elections: Candidate Selection and Success.” *The Journal of Politics* 65(3):833–50. doi: 10.1111/1468-2508.00214.

Franklin, Jennifer. 2001. “Interpreting the Numbers: Using a Narrative to Help Others Read Student Evaluations of Your Teaching Accurately.” *New Directions for Teaching and Learning* 2001(87):85–100. doi: 10.1002/tl.10001.

Garcia-Retamero, Rocio, and Esther López-Zafra. 2006. “Prejudice against Women in Male-Congenial Environments: Perceptions of Gender Role Congruity in Leadership.” *Sex Roles* 55(1–2):51–61. doi: 10.1007/s11199-006-9068-1.

Heilman, Madeline E. 2012a. “Gender Stereotypes and Workplace Bias.” *Research in Organizational Behavior* 32:113–35. doi: 10.1016/j.riob.2012.11.003.

Heilman, Madeline E. 2012b. “Gender Stereotypes and Workplace Bias.” *Research in Organizational Behavior* 32:113–35. doi: 10.1016/j.riob.2012.11.003.

Hitsch, Günter J., Ali Hortaçsu, and Dan Ariely. 2010. “What Makes You Click?—Mate Preferences in Online Dating.” *Quantitative Marketing and Economics* 8(4):393–427. doi: 10.1007/s11129-010-9088-6.

Jacobs, Struan. 2018a. “How Role Replaced Personality as a Major Category of Sociology.” *The American Sociologist; Washington* 49(2):280–98. doi: http://dx.doi.org.www.libproxy.wvu.edu/10.1007/s12108-017-9354-0.

Jacobs, Struan. 2018b. “How Role Replaced Personality as a Major Category of Sociology.” *The American Sociologist* 49(2):280–98. doi: 10.1007/s12108-017-9354-0.

Johnson, Stefanie K., Susan Elaine Murphy, Selamawit Zewdie, and Rebecca J. Reichard. 2008. “The Strong, Sensitive Type: Effects of Gender Stereotypes and Leadership Prototypes on the Evaluation of Male and Female Leaders.” *Organizational Behavior and Human Decision Processes* 106(1):39–60. doi: 10.1016/j.obhdp.2007.12.002.

Kalender, İlker. 2015. "Measurement Invariance of Student Evaluation of Teaching across Groups Defined by Course-Related Variables." *International Online Journal of Educational Sciences*. doi: 10.15345/iojes.2015.04.006.

Kay, Aaron C., Danielle Gaucher, Jennifer M. Peach, Kristin Laurin, Justin Friesen, Mark P. Zanna, and Steven J. Spencer. 2009. "Inequality, discrimination, and the power of the status quo: Direct evidence for a motivation to see the way things are as the way they should be." *Journal of Personality and Social Psychology* 97(3):421–34. doi: 10.1037/a0015997.

Kline, Rex B. 2015. *Principles and Practice of Structural Equation Modeling, Fourth Edition*. New York, UNITED STATES: Guilford Publications.

Kobrynowicz, Diane, and Monica Biernat. 1997. "Decoding Subjective Evaluations: How Stereotypes Provide Shifting Standards." *Journal of Experimental Social Psychology* 33(6):579–601. doi: 10.1006/jesp.1997.1338.

Koenig, Anne M., and Alice H. Eagly. 2014. "Evidence for the Social Role Theory of Stereotype Content: Observations of Groups' Roles Shape Stereotypes." *Journal of Personality and Social Psychology* 107(3):371–92. doi: 10.1037/a0037215.

Liden, Robert C., Dean Stilwell, and Gerald R. Ferris. 1996. "The Effects of Supervisor and Subordinate Age on Objective Performance and Subjective Performance Ratings." *Human Relations; Thousand Oaks* 49(3):327.

Livingston, Robert W., Ashleigh Shelby Rosette, and Ella F. Washington. 2012. "Can an Agentic Black Woman Get Ahead? The Impact of Race and Interpersonal Dominance on Perceptions of Female Leaders." *Psychological Science* 23(4):354–58. doi: 10.1177/0956797611428079.

Lynch, Karen Danna. 2007. "Modeling Role Enactment: Linking Role Theory and Social Cognition." *Journal for the Theory of Social* B*ehaviour* 37(4):379–99.

MacNell, Lillian, Adam Driscoll, and Andrea N. Hunt. 2015a. "What's in a Name: Exposing Gender Bias in Student Ratings of Teaching." *Innovative Higher Education* 40(4):291–303. doi: 10.1007/s10755-014-9313-4.

MacNell, Lillian, Adam Driscoll, and Andrea N. Hunt. 2015b. "What's in a Name: Exposing Gender Bias in Student Ratings of Teaching." *Innovative Higher Education* 40(4):291–303. doi: 10.1007/s10755-014-9313-4.

McMurtrie, Beth. 2019. "'Brilliant' Philosophers and 'Funny' Psychology Instructors: What a Data-Visualization Tool Tells Us About How Students See Their Professors." *The Chronicle of Higher Education*, November 14.

Mitchell, Kristina M. W., and Jonathan Martin. 2018. "Gender Bias in Student Evaluations." *PS: Political Science & Politics* 51(3):648–52. doi: 10.1017/S104909651800001X.

Moss-Racusin, Corinne A., Julie E. Phelan, and Laurie A. Rudman. 2010. "When Men Break the Gender Rules: Status Incongruity and Backlash against Modest Men." *Psychology of Men & Masculinity* 11(2):140–51. doi: 10.1037/a0018093.

National Center for Science and Engineering Statistics. 2021. *Doctorate Recipients, by Sex and Major Field of Study: 2020*. Alexandria, VA.

Reid, Landon D. 2010. "The Role of Perceived Race and Gender in the Evaluation of College Teaching on RateMyProfessors.Com." *Journal of Diversity in Higher Education* 3(3):137–52. doi: 10.1037/a0019865.

Ridgeway, Cecilia L., and Shelley J. Correll. 2004. "Unpacking the Gender System: A Theoretical Perspective on Gender Beliefs and Social Relations." *Gender & Society* 18(4):510–31. doi: 10.1177/0891243204265269.

Ridgeway, Cecilia L., and Lynn Smith-Lovin. 1999. "THE GENDER SYSTEM AND INTERACTION." *Annual Review of Sociology* 25(1):191–216. doi: 10.1146/annurev.soc.25.1.191.

Rudman, Laurie A., and Peter Glick. 2001a. "Prescriptive Gender Stereotypes and Backlash Toward Agentic Women." *Journal of Social Issues* 57(4):743–62. doi: 10.1111/0022-4537.00239.

Rudman, Laurie A., and Peter Glick. 2001b. "Prescriptive Gender Stereotypes and Backlash Toward Agentic Women." *Journal of Social Issues* 57(4):743–62. doi: 10.1111/0022-4537.00239.

Rudman, Laurie A., Corinne A. Moss-Racusin, Julie E. Phelan, and Sanne Nauts. 2011. "Status Incongruity and Backlash Effects: Defending the Gender Hierarchy Motivates Prejudice against Female Leaders." *Journal of Experimental Social Psychology* 48(1):165–79. doi: 10.1016/j.jesp.2011.10.008.

Simpson, Ruth. 2004a. "Masculinity at Work: The Experiences of Men in Female Dominated Occupations." *Work, Employment & Society* 18(2):349–68.

Simpson, Ruth. 2004b. "Masculinity at Work: The Experiences of Men in Female Dominated Occupations." *Work, Employment and Society* 18(2):349–68. doi: 10.1177/09500172004042773.

Smith, Bettye P., and Billy Hawkins. 2011. "Examining Student Evaluations of Black College Faculty: Does Race Matter?" *The Journal of Negro Education* 80(2):149–62.

Smith, D. Randall, Nancy DiTomaso, George F. Farris, and Rene Cordero. 2001. "Favoritism, Bias, and Error in Performance Ratings of Scientists and Engineers: The Effects of Power, Status, and Numbers." *Sex Roles; New York* 45(5/6):337–58. doi: http://dx.doi.org.www.libproxy.wvu.edu/10.1023/A:1014309631243.

Smith, David G., Judith E. Rosensetin, Margaret C. Nikolov, and Darby A. Chaney. 2019. "The Power of Language: Gender, Status, and Agency in Performance Evaluations." *Sex Roles; New York* 80(3–4):159. doi: http://dx.doi.org.www.libproxy.wvu.edu/10.1007/s11199-018-0923-7.

Sprague, Joey, and Kelley Massoni. 2005. "Student Evaluations and Gendered Expectations: What We Can't Count Can Hurt Us." *Sex Roles* 53(11–12):779–93. doi: 10.1007/s11199-005-8292-4.

Wachtel, Howard K. 1998. "Student Evaluation of College Teaching Effectiveness: A Brief Review." *Assessment & Evaluation in Higher Education* 23(2):191–212. doi: 10.1080/0260293980230207.

Williams, Christine L. 1992. "The Glass Escalator: Hidden Advantages for Men in the Female Professions Four Papers on Inequality." *Social Problems* 39(3):253–67.

Wingfield, Adia Harvey. 2009. "Racializing the Glass Escalator: Reconsidering Men's Experiences with Women's Work." *Gender & Society* 23(1):5–26. doi: 10.1177/0891243208323054.

Wood, Wendy, and Alice H. Eagly. 2002. "A Cross-Cultural Analysis of the Behavior of Women and Men: Implications for the Origins of Sex Differences." *Psychological Bulletin* 128(5):699–727. doi: 10.1037/0033-2909.128.5.699.

APPENDICES

Appendix 1: Categorization of Disciplines by Gender Dominance

Subject	Gender Dominance Category
Accounting	Neutral
Advertising	Woman-Dominated
African/American Studies	Woman-Dominated
Agriculture Forestry & Consumer Sci	Woman-Dominated
Agricultural Biochemistry	Woman-Dominated
Agriculture & Extension Edu	Woman-Dominated
Agriculture & Resource Econ	Woman-Dominated
Agronomy	Man-Dominated
Animal Nutrition	Woman-Dominated
Animal Physiology	Woman-Dominated
Animal Production	Woman-Dominated
Animal and Veterinary Science	Woman-Dominated
Applied & Environment Microbiology	Woman-Dominated
Arabic	Woman-Dominated
Art	Woman-Dominated
Art History	Woman-Dominated
Arts and Sciences	Neutral
Astronomy	Man-Dominated
Athletic Coaching Education	Woman-Dominated
Athletic Training	Woman-Dominated
Biology	Woman-Dominated
Biomedical Engineering	Man-Dominated
Biometric Systems	Man-Dominated
Business Administration (BUSA)	Neutral
Business Core	Neutral
Business Law	Neutral
Chemical Engineering	Man-Dominated
Chemistry	Man-Dominated
Child Dev / Family Studies	Woman-Dominated
Chinese	Woman-Dominated
Civil Engineering	Man-Dominated
Classics	Woman-Dominated
Comm Sciences and Disorders	Woman-Dominated
Communication Studies	Woman-Dominated
Computer Engineering	Man-Dominated
Computer Science	Man-Dominated
Counseling	Woman-Dominated

Curriculum and Instruction	Woman-Dominated
Dance	Woman-Dominated
Design Studies	Woman-Dominated
Design and Merchandising	Woman-Dominated
Disability Studies	Woman-Dominated
Economics	Neutral
Education	Woman-Dominated
Education and Human Services	Woman-Dominated
Educational Psychology	Woman-Dominated
Electrical Engineering	Man-Dominated
Energy Land Management	Man-Dominated
Engineering	Man-Dominated
English	Woman-Dominated
English as a Second Language	Woman-Dominated
Entomology	Man-Dominated
Entrepreneurship	Neutral
Environmental Protection	Neutral
Fashion Dress & Merchandising	Woman-Dominated
Film	Woman-Dominated
Finance	Neutral
Food Science & Technology	Man-Dominated
Foreign Culture	Woman-Dominated
Foreign Lit in Translation	Woman-Dominated
Forensic and Investigative Science	Neutral
Forest Hydrology	Man-Dominated
Forest Management	Man-Dominated
Forestry	Man-Dominated
French	Woman-Dominated
Genetics	Man-Dominated
Geography	Man-Dominated
Geology	Man-Dominated
German	Woman-Dominated
Gerontology	Woman-Dominated
Global Supply Chain Management	Man-Dominated
History	Man-Dominated
Honors	Neutral
Horticulture	Man-Dominated
Hospitality/Tourism	Woman-Dominated
Human Nutrition and Foods	Woman-Dominated
Human Resource Management	Woman-Dominated
Humanities	Woman-Dominated

Industrial Engineering	Man-Dominated
Industrial Hygiene & Safety	Man-Dominated
Interior Design	Woman-Dominated
International Business	Neutral
International Studies	Woman-Dominated
Italian	Woman-Dominated
Japanese	Woman-Dominated
Journalism	Woman-Dominated
Landscape Architecture	Neutral
Language Teaching Methods	Woman-Dominated
Leadership Studies	Woman-Dominated
Linguistics	Woman-Dominated
Management	Neutral
Management Information Systems	Neutral
Marketing	Neutral
Mathematics	Man-Dominated
Mechanical and Aerospace Engr	Man-Dominated
Mining Engineering	Man-Dominated
Multidisciplinary Studies	Woman-Dominated
Music	Woman-Dominated
Native American Studies	Woman-Dominated
Petroleum and Natural Gas Engr	Man-Dominated
Philosophy	Man-Dominated
Physical Act / Sport Sciences	Man-Dominated
Physical Education	Woman-Dominated
Physical Education/Teaching	Woman-Dominated
Physics	Man-Dominated
Plant Pathology	Man-Dominated
Plant Science	Neutral
Political Science	Neutral
Psychology	Woman-Dominated
Public Relations	Woman-Dominated
Reading	Woman-Dominated
Recreation Parks & Tourism Res	Neutral
Religious Studies	Man-Dominated
Resource Management	Neutral
Russian	Woman-Dominated
Safety Management	Man-Dominated
Slavic & Eastern European St	Woman-Dominated
Social Work	Woman-Dominated
Sociology and Anthropology	Neutral

Spanish	Woman-Dominated
Special Education	Woman-Dominated
Sport Management	Man-Dominated
Sport and Exercise Psychology	Woman-Dominated
Statistics	Man-Dominated
Strategic Communications	Woman-Dominated
Theatre	Woman-Dominated
UTeach Program	Woman-Dominated
Veterinary Science	Woman-Dominated
Wildlife and Fisheries Management	Neutral
Women and Gender Studies	Woman-Dominated
Wood Science	Man-Dominated

Appendix 2: Gender Role Congruent Partial Invariant Loadings Model Means Comparison: White Women Reference Group

	Content-Related-Assignments Constrained							
	White Men (*N*=23,870)		White Women (*N*=17,256)		Racially/ethnically Minoritized Men (*N*=7,925)		Racially/ethnically Minoritized Women (*N*=3,332)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	a	0.84***	a	0.85***	a	0.85***	a	0.87***
Content-Thought-Provoking	1.05***	0.80***	1.11***	0.83***	1.05***	0.80***	1.00***	0.82***
Material-Useful	1.15***	0.90***	1.19***	0.90***	1.13***	0.89***	1.16***	0.92***
Overall-Learning	0.49***	0.32***	0.41***	0.27***	0.43***	0.29***	0.30***	0.21***
Instructor								
Positive-Learning-Environment	0.96	0.85***	0.93	0.85***	0.96	0.86***	0.89	0.85***
Instructor-Organized	0.98	0.82***	1.00	0.82***	0.96	0.82***	1.01	0.86***
Instructor-Feedback	1.14	0.88***	1.14	0.87***	1.14	0.88***	1.14	0.90***
Overall-Learning	0.69	0.50***	0.72	0.49***	0.74	0.54***	0.84	0.62***
Mean Overall	-0.12***	-0.16***	a	a	-0.18***	-0.22***	-0.22***	-0.25***
Mean Instructor	-0.24	-0.28***	a	a	-0.33	-0.37***	-0.30	-0.33***
R^2	0.978		0.981		0.979		0.984	
$\chi 2$	*df*=62, 1314.34***							
CFI	0.995							
RMSEA	0.039							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Appendix 3: Gender Role Congruent Partial Invariant Loadings Model Means Comparison: Racially/ethnically Minoritized Men Reference Group

	Content-Related-Assignments Constrained							
	White Men (*N*=23,870)		White Women (*N*=17,256)		Racially/ethnically Minoritized Men (*N*=7,925)		Racially/ethnically Minoritized Women (*N*=3,332)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	a	0.84***	a	0.85***	a	0.85***	a	0.87***
Content-Thought-Provoking	1.05***	0.80***	1.11***	0.83***	1.03***	0.80***	1.00***	0.82***
Material-Useful	1.15***	0.90***	1.18***	0.90***	1.13***	0.89***	1.16***	0.92***
Overall-Learning	0.49***	0.32***	0.39***	0.26***	0.42***	0.28***	0.32***	0.22***
Instructor								
Positive-Learning-Environment	0.61	0.85***	0.59	0.85***	0.62	0.86***	0.58	0.85***
Instructor-Organized	0.63	0.82***	0.63	0.82***	0.62	0.82***	0.66	0.86***
Instructor-Feedback	0.73	0.88***	0.73	0.87***	0.73	0.88***	0.73	0.90***
Overall-Learning	0.44	0.50***	0.47	0.50***	0.48	0.56***	0.53	0.61***
Mean Overall	0.06***	0.08***	0.17***	0.24***	a	a	-0.04***	-0.05
Mean Instructor	0.13	0.10***	0.51	0.45***	a	a	0.04*	0.03*
R^2	0.979		0.981		0.980		0.984	
$\chi 2$	*df*=62, 1330.77***							
CFI	0.995							
RMSEA	0.040							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Appendix 4: Gender Role Congruent Partial Invariant Loadings Model Means Comparison: Racially/ethnically Minoritized Women Reference Group

	Content-Related-Assignments Constrained							
	White Men (*N*=23,870)		White Women (*N*=17,256)		Racially/ethnically Minoritized Men (*N*=7,925)		Racially/ethnically Minoritized Women (*N*=3,332)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	a	0.84***	a	0.85***	a	0.85***	a	0.87***
Content-Thought-Provoking	1.04***	0.79***	1.11***	0.83***	1.03***	0.80***	1.00***	0.82***
Material-Useful	1.16***	0.90***	1.18***	0.90***	1.13***	0.90***	1.16***	0.92***
Overall-Learning	0.49***	0.32***	0.40***	0.27***	0.42***	0.28***	0.30***	0.21***
Instructor								
Positive-Learning-Environment	0.55	0.85***	0.53	0.85***	0.55	0.86***	0.51	0.85***
Instructor-Organized	0.56	0.82***	0.57	0.82***	0.55	0.82***	0.58	0.86***
Instructor-Feedback	0.65	0.88***	0.65	0.87***	0.65	0.88***	0.64	0.90***
Overall-Learning	0.39	0.50***	0.42	0.50***	0.43	0.56***	0.48	0.62***
Mean Overall	0.10***	0.13***	0.21***	0.29***	0.04*	0.05*	a	a
Mean Instructor	0.09	0.06**	0.51	0.41***	-0.06	-0.04	a	a
R^2	0.979		0.981		0.980		0.984	
$\chi 2$	*df*=62, 1326.67***							
CFI	0.995							
RMSEA	0.039							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Appendix 5: Gender Role Congruent Partial Invariant Loadings Model Means Comparison: White Women Reference Group

	Content-Thought-Provoking Constrained							
	White Men (*N*=23,870)		White Women (*N*=17,256)		Racially/ethnically Minoritized Men (*N*=7,925)		Racially/ethnically Minoritized Women (*N*=3,332)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.95***	0.84***	0.90***	0.85***	0.96***	0.85***	1.00***	0.87***
Content-Thought-Provoking	a	0.80***	a	0.83***	a	0.80***	a	0.82***
Material-Useful	1.10***	0.90***	1.08***	0.90***	1.08***	0.89***	1.16***	0.92***
Overall-Learning	0.46***	0.32***	0.37***	0.27***	0.41***	0.29***	0.30***	0.21***
Instructor								
Positive-Learning-Environment	0.83	0.85***	0.80	0.85***	0.84	0.86***	0.77	0.85***
Instructor-Organized	0.85	0.82***	0.87	0.82***	0.83	0.82***	0.88	0.86***
Instructor-Feedback	0.99	0.88***	0.99	0.87***	0.99	0.88***	0.99	0.90***
Overall-Learning	0.60	0.51***	0.62	0.49***	0.65	0.55***	0.73	0.62***
Mean Overall	-0.12***	-0.16***	a	a	-0.18***	-0.22***	-0.22***	-0.25***
Mean Instructor	-0.27	-0.28***	a	a	-0.38	-0.37***	-0.34	-0.33***
R^2	0.978		0.981		0.979		0.984	
$\chi 2$	*df*=62, 1314.34***							
CFI	0.995							
RMSEA	0.039							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Appendix 6: Gender Role Congruent Partial Invariant Loadings Model Means Comparison: Racially/ethnically Minoritized Men Reference Group

	Content-Thought-Provoking Constrained							
	White Men (*N*=23,870)		White Women (*N*=17,256)		Racially/ethnically Minoritized Men (*N*=7,925)		Racially/ethnically Minoritized Women (*N*=3,332)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.96***	0.84***	0.90***	0.85***	0.97***	0.85***	1.00***	0.87***
Content-Thought-Provoking	a	0.79***	a	0.83***	a	0.80***	a	0.82***
Material-Useful	1.11***	0.90***	1.06***	0.90***	0.97***	0.89***	1.16***	0.92***
Overall-Learning	0.47***	0.32***	0.35***	0.26***	0.41***	0.28***	0.32***	0.22***
Instructor								
Positive-Learning-Environment	0.82	0.85***	0.79	0.85***	0.83	0.86***	0.77	0.85***
Instructor-Organized	0.84	0.82***	0.85	0.82***	0.83	0.82***	0.88	0.86***
Instructor-Feedback	0.97	0.88***	0.97	0.87***	0.97	0.88***	0.97	0.90***
Overall-Learning	0.58	0.50***	0.63	0.50***	0.64	0.56***	0.70	0.61***
Mean Overall	0.06***	0.08***	0.19***	0.24***	a	a	-0.04**	-0.05*
Mean Instructor	0.10	0.10***	0.38	0.45***	a	a	0.03	0.03
R^2	0.979		0.981		0.980		0.984	
$\chi 2$	*df*=62, 1330.77***							
CFI	0.995							
RMSEA	0.040							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Appendix 7: Gender Role Congruent Partial Invariant Loadings Model Means Comparison: Racially/ethnically Minoritized Women Reference Group

	Content-Thought-Provoking Constrained							
	White Men (*N*=23,870)		White Women (*N*=17,256)		Racially/ethnically Minoritized Men (*N*=7,925)		Racially/ethnically Minoritized Women (*N*=3,332)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.96***	0.84***	0.90***	0.85***	0.97***	0.85***	1.00***	0.87***
Content-Thought-Provoking	a	0.79***	a	0.83***	a	0.80***	a	0.82***
Material-Useful	1.11***	0.90***	1.06***	0.90***	1.09***	0.90***	1.16***	0.92***
Overall-Learning	0.47***	0.32***	0.36***	0.27***	0.41***	0.28***	0.30***	0.21***
Instructor								
Positive-Learning-Environment	0.63	0.85***	0.61	0.85***	0.64	0.86***	0.60	0.85***
Instructor-Organized	0.65	0.82***	0.66	0.82***	0.64	0.82***	0.68	0.86***
Instructor-Feedback	0.75	0.88***	0.75	0.87***	0.75	0.88***	0.75	0.90***
Overall-Learning	0.45	0.50***	0.48	0.50***	0.50	0.56***	0.55	0.62***
Mean Overall	0.10***	0.13***	0.23***	0.29***	0.04*	0.05*	a	a
Mean Instructor	0.08	0.06**	0.44	0.41***	-0.05	-0.04	a	a
R^2	0.979		0.981		0.980		0.984	
$\chi 2$	*df*=62, 1326.67***							
CFI	0.995							
RMSEA	0.039							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Appendix 8: Gender Role Incongruent Partial Invariant Loadings Model Means Comparison: White Women Reference Group

	White Women as Reference Group							
	White Men (*N*=11,108)		White Women (*N*=9,952)		Racially/ethnically Minoritized Men (*N*=2,364)		Racially/ethnically Minoritized Women (*N*=2,674)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.60	0.86***	0.60	0.85***	0.60	0.84***	0.56	0.83***
Content-Thought-Provoking	0.64	0.83***	0.61	0.79***	0.63	0.80***	0.66	0.83***
Material-Useful	0.68	0.90***	0.68	0.90***	0.68	0.89***	0.68	0.90***
Overall-Learning	0.24	0.29***	0.27	0.30***	0.25	0.27***	0.28	0.30***
Instructor								
Positive-Learning-Environment	0.83	0.85***	0.83	0.87***	0.83	0.84***	0.83	0.86***
Instructor-Organized	0.86	0.82***	0.86	0.84***	0.86	0.82***	0.86	0.83***
Instructor-Feedback	0.99	0.87***	0.99	0.89***	0.99	0.84***	0.99	0.89***
Overall-Learning	0.62	0.50***	0.62	0.50***	0.62	0.49***	0.62	0.52***
Mean Overall	0.16	0.13***	a	a	0.27	0.24***	-0.06	-0.05***
Mean Instructor	0.10	0.12***	a	a	0.16	0.20***	-0.16	-0.16***
R^2	0.981		0.981		0.977		0.982	
$\chi 2$	*df*=70, 1057.13***							
CFI	0.992							
RMSEA	0.046							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Appendix 9: Gender Role Incongruent Partial Invariant Loadings Model Means Comparison: Racially/ethnically Minoritized Men Reference Group

	Racially/ethnically Minoritized Men as Reference Group							
	White Men (*N*=11,108)		White Women (*N*=9,952)		Racially/ethnically Minoritized Men (*N*=2,364)		Racially/ethnically Minoritized Women (*N*=2,674)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.58	0.86***	0.58	0.85***	0.58	0.84***	0.54	0.82***
Content-Thought-Provoking	0.61	0.83***	0.61	0.79***	0.61	0.80***	0.65	0.83***
Material-Useful	0.67	0.90***	0.67	0.90***	0.67	0.89***	0.67	0.90***
Overall-Learning	0.22	0.27***	0.28	0.31***	0.24	0.27***	0.28	0.31***
Instructor								
Positive-Learning-Environment	0.72	0.85***	0.72	0.87***	0.72	0.84***	0.72	0.86***
Instructor-Organized	0.74	0.82***	0.74	0.84***	0.74	0.82***	0.74	0.83***
Instructor-Feedback	0.85	0.87***	0.85	0.88***	0.85	0.84***	0.85	0.89***
Overall-Learning	0.53	0.50***	0.53	0.49***	0.53	0.49***	0.53	0.52***
Mean Overall	-0.17	-0.13***	-0.34	-0.26***	a	a	-0.40	-0.31***
Mean Instructor	-0.10	-0.10***	-0.22	-0.21***	a	a	-0.40	-0.35***
R^2	0.981		0.981		0.977		0.982	
χ^2	*df*=70, 1015.76***							
CFI	0.993							
RMSEA	0.046							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Appendix 10: Gender Role Incongruent Partial Invariant Loadings Model Means Comparison: Racially/ethnically Minoritized Women Reference Group

	Racially/ethnically Minoritized Women as Reference Group							
	White Men (*N*=11,108)		White Women (*N*=9,952)		Racially/ethnically Minoritized Men (*N*=2,364)		Racially/ethnically Minoritized Women (*N*=2,674)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.63	0.86***	0.63	0.85***	0.63	0.84***	0.59	0.83***
Content-Thought-Provoking	0.67	0.83***	0.64	0.79***	0.67	0.80***	0.69	0.83***
Material-Useful	0.72	0.90***	0.72	0.90***	0.72	0.89***	0.72	0.90***
Overall-Learning	0.26	0.29***	0.28	0.30***	0.26	0.27***	0.29	0.30***
Instructor								
Positive-Learning-Environment	0.56	0.85***	0.56	0.87***	0.56	0.84***	0.56	0.86***
Instructor-Organized	0.57	0.82***	0.57	0.84***	0.57	0.82***	0.57	0.83***
Instructor-Feedback	0.55	0.87***	0.66	0.89***	0.66	0.84***	0.66	0.89***
Overall-Learning	0.41	0.50***	0.41	0.50***	0.41	0.49***	0.41	0.52***
Mean Overall	0.19	0.16***	0.03	0.02	0.29	0.27***	a	a
Mean Instructor	0.38	0.29***	0.21	0.16***	0.46	0.38***	a	a
R^2	0.980		0.981		0.977		0.982	
$\chi 2$	*df*=70, 1059.93***							
CFI	0.992							
RMSEA	0.047							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Appendix 11: Gender Role Neutral Partial Invariant Loadings Model Means Comparison: White Women Reference Group

	Overall-Learning on Instructor Constrained							
	White Men (*N*=4,889)		White Women (*N*=6,018)		Racially/ethnically Minoritized Men (*N*=944)		Racially/ethnically Minoritized Women (*N*=791)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.94	0.85***	0.90	0.85***	0.97	0.80***	0.92	0.83***
Content-Thought-Provoking	0.98	0.82***	0.98	0.81***	0.88	0.72***	0.98	0.82***
Material-Useful	1.07	0.90***	1.07	0.91***	1.07	0.86***	1.07	0.88***
Overall-Learning	0.37	0.26***	0.46	0.33***	0.65	0.42***	0.41	0.27***
Instructor								
Positive-Learning-Environment	1.14***	0.84***	1.36***	0.85***	1.57***	0.82***	1.03***	0.82***
Instructor-Organized	1.27***	0.81***	1.36***	0.81***	1.34***	0.74***	1.19***	0.82***
Instructor-Feedback	1.40***	0.88***	1.64***	0.90***	1.90***	0.87***	1.32***	0.89***
Overall-Learning	a	0.56***	a	0.53***	a	0.43***	a	0.62***
Mean Overall	0.07	0.09***	a	a	-0.23	-0.28***	-0.21	-0.23***
Mean Instructor	0.10***	0.16***	a	a	-0.19***	-0.36***	-0.25***	-0.30***
R^2	0.977		0.982		0.967		0.974	
χ^2	*df*=62, 511.82***							
CFI	0.993							
RMSEA	0.048							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Appendix 12: Gender Role Neutral Partial Invariant Loadings Model Means Comparison: Racially/ethnically Minoritized Men Reference Group

	Overall-Learning on Instructor Constrained							
	White Men (*N*=4,889)		White Women (*N*=6,018)		Racially/ethnically Minoritized Men (*N*=944)		Racially/ethnically Minoritized Women (*N*=791)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.84	0.85***	0.81	0.85***	0.86	0.80***	0.82	0.83***
Content-Thought-Provoking	0.88	0.82***	0.88	0.81***	0.79	0.73***	0.88	0.82***
Material-Useful	0.96	0.90***	0.96	0.91***	0.96	0.87***	0.96	0.88***
Overall-Learning	0.33	0.26***	0.41	0.32***	0.60	0.45***	0.39	0.29***
Instructor								
Positive-Learning-Environment	1.14***	0.84***	1.35***	0.85***	1.71***	0.83***	1.07***	0.83***
Instructor-Organized	1.25***	0.81***	1.36***	0.82***	1.51***	0.75***	1.22***	0.82***
Instructor-Feedback	1.40***	0.88***	1.62***	0.90***	2.07***	0.87***	1.36***	0.89***
Overall-Learning	a	0.56***	a	0.54***	a	0.40***	a	0.61***
Mean Overall	0.23	0.25***	0.13	0.14***	a	a	-0.09	-0.09
Mean Instructor	0.27***	0.42***	0.13***	0.21***	a	a	-0.08	-0.09
R^2	0.977		0.982		0.967		0.974	
χ^2	*df*=62, 573.88***							
CFI	0.992							
RMSEA	0.051							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Appendix 13: Gender Role Neutral Partial Invariant Loadings Model Means Comparison: Racially/ethnically Minoritized Women Reference Group

	Overall-Learning on Instructor Constrained							
	White Men (*N*=4,889)		White Women (*N*=6,018)		Racially/ethnically Minoritized Men (*N*=944)		Racially/ethnically Minoritized Women (*N*=791)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.85	0.85***	0.82	0.85***	0.88	0.80***	0.83	0.84***
Content-Thought-Provoking	0.89	0.82***	0.89	0.81***	0.80	0.72***	0.89	0.83***
Material-Useful	0.97	0.90***	0.97	0.91***	0.97	0.86***	0.97	0.89***
Overall-Learning	0.33	0.26***	0.41	0.32***	0.62	0.45***	0.40	0.29***
Instructor								
Positive-Learning-Environment	1.14***	0.84***	1.35***	0.85***	1.69***	0.82***	1.06***	0.83***
Instructor-Organized	1.26***	0.81***	1.36***	0.82***	1.46***	0.74***	1.23***	0.83***
Instructor-Feedback	1.40***	0.88***	1.63***	0.90***	2.05***	0.87***	1.36***	0.89***
Overall-Learning	a	0.56***	a	0.53***	a	0.40***	a	0.60***
Mean Overall	0.14	0.16***	0.04	0.04	-0.19	-0.21***	a	a
Mean Instructor	0.20***	0.31***	0.07**	0.11**	-0.11***	-0.22***	a	a
R^2	0.977		0.982		0.967		0.975	
χ^2	*df*=62, 561.83***							
CFI	0.992							
RMSEA	0.051							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Appendix 14: Gender Role Neutral Partial Invariant Loadings Model Means Comparison: White Women Reference Group Positive-Learning-Environment Constrained

	Positive-Learning-Environment on Instructor Constrained							
	White Men (*N*=4,889)		White Women (*N*=6,018)		Racially/ethnically Minoritized Men (*N*=944)		Racially/ethnically Minoritized Women (*N*=791)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.61	0.85***	0.59	0.85***	0.63	0.80***	0.60	0.83***
Content-Thought-Provoking	0.63	0.82***	0.64	0.81***	0.57	0.72***	0.64	0.82***
Material-Useful	0.69	0.90***	0.69	0.91***	0.69	0.86***	0.69	0.88***
Overall-Learning	0.24	0.26***	0.30	0.33***	0.42	0.42***	0.27	0.27***
Instructor								
Positive-Learning-Environment	a	0.84***	a	0.85***	a	0.82***	a	0.82***
Instructor-Organized	1.12***	0.81***	1.00***	0.81***	0.85***	0.74***	1.16***	0.82***
Instructor-Feedback	1.23***	0.88***	1.20***	0.90***	1.21***	0.87***	1.28***	0.89***
Overall-Learning	0.88***	0.56***	0.73***	0.53***	0.64***	0.43***	0.98***	0.62***
Mean Overall	0.11	0.09***	a	a	-0.35	-0.28***	-0.32	-0.23***
Mean Instructor	0.12***	0.16***	a	a	-0.31***	-0.36***	-0.26***	-0.30***
R^2	0.977		0.982		0.967		0.974	
χ^2	*df*=62, 511.82***							
CFI	0.993							
RMSEA	0.048							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Appendix 15: Gender Role Neutral Partial Invariant Loadings Model Means Comparison: Racially/ethnically Minoritized Men Reference Group Positive-Learning-Environment Constrained

	Positive-Learning-Environment on Instructor Constrained							
	White Men (*N*=4,889)		White Women (*N*=6,018)		Racially/ethnically Minoritized Men (*N*=944)		Racially/ethnically Minoritized Women (*N*=791)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.59	0.85***	0.57	0.85***	0.61	0.80***	0.58	0.83***
Content-Thought-Provoking	0.62	0.82***	0.62	0.81***	0.56	0.73***	0.62	0.82***
Material-Useful	0.67	0.90***	0.67	0.91***	0.67	0.87***	0.67	0.88***
Overall-Learning	0.23	0.26***	0.29	0.32***	0.43	0.45***	0.27	0.29***
Instructor								
Positive-Learning-Environment	a	0.84***	a	0.85***	a	0.83***	a	0.83***
Instructor-Organized	1.09***	0.81***	1.01***	0.82***	0.88***	0.75***	1.15***	0.82***
Instructor-Feedback	1.22***	0.88***	1.21***	0.90***	1.21***	0.87***	1.28***	0.89***
Overall-Learning	0.87***	0.56***	0.74***	0.54***	0.58***	0.40***	0.94***	0.60***
Mean Overall	0.32	0.25***	0.18	0.14***	a	a	-0.12	-0.09
Mean Instructor	0.31***	0.42***	0.18***	0.21***	a	a	-0.08	-0.09
R^2	0.977		0.982		0.967		0.974	
χ^2	*df*=62, 573.88***							
CFI	0.992							
RMSEA	0.051							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*$p<0.05$; **$p<0.01$; ***$p\leq0.001$

Appendix 16: Gender Role Neutral Partial Invariant Loadings Model Means Comparison: Racially/ethnically Minoritized Women Reference Group Positive-Learning-Environment Constrained

	Positive-Learning-Environment on Instructor Constrained							
	White Men (*N*=4,889)		White Women (*N*=6,018)		Racially/ethnically Minoritized Men (*N*=944)		Racially/ethnically Minoritized Women (*N*=791)	
	B	*β*	*B*	*β*	*B*	*β*	*B*	*β*
Overall								
Content-Related-Assignments	0.67	0.85***	0.65	0.85***	0.69	0.80***	0.66	0.84***
Content-Thought-Provoking	0.70	0.82***	0.70	0.81***	0.63	0.72***	0.70	0.83***
Material-Useful	0.76	0.90***	0.76	0.91***	0.76	0.86***	0.76	0.89***
Overall-Learning	0.26	0.26***	0.32	0.32***	0.49	0.45***	0.31	0.29***
Instructor								
Positive-Learning-Environment	a	0.84***	a	0.85***	a	0.82***	a	0.83***
Instructor-Organized	1.10***	0.81***	1.00***	0.82***	0.86***	0.74***	1.15***	0.82***
Instructor-Feedback	1.23***	0.88***	1.20***	0.90***	1.21***	0.87***	1.28***	0.89***
Overall-Learning	0.88***	0.56***	0.74***	0.53***	0.59***	0.40***	0.94***	0.60***
Mean Overall	0.17	0.16***	0.05	0.04	-0.25	-0.21***	a	a
Mean Instructor	0.22***	0.31***	0.09**	0.11**	-0.19***	-0.22***	a	a
R^2	0.977		0.982		0.967		0.975	
χ^2	*df*=62, 561.83***							
CFI	0.992							
RMSEA	0.051							

B=unstandardized, *β*=standardized
a Not reported because of constraints
*p<0.05; **p<0.01; ***$p\leq$0.001

Appendix 17: Description of the 48 Instructor Groups[17]

Gender	Race/ethnicity	Gender Dominance of Discipline	Level of Congruity	Time Period	Question
Man	White	Man-Dominated	Congruent	Old	1: Course
Man	White	Women-Dominated	Incongruent	Old	1: Course
Man	White	Neutral	Neutral	Old	1: Course
Man	Racially/ethnically minoritized	Man-Dominated	Congruent	Old	1: Course
Man	Racially/ethnically minoritized	Women-Dominated	Incongruent	Old	1: Course
Man	Racially/ethnically minoritized	Neutral	Neutral	Old	1: Course
Woman	White	Man-Dominated	Incongruent	Old	1: Course
Woman	White	Women-Dominated	Congruent	Old	1: Course
Woman	White	Neutral	Neutral	Old	1: Course
Woman	Racially/ethnically minoritized	Man-Dominated	Incongruent	Old	1: Course
Woman	Racially/ethnically minoritized	Women-Dominated	Congruent	Old	1: Course
Woman	Racially/ethnically minoritized	Neutral	Neutral	Old	1: Course
Man	White	Man-Dominated	Congruent	Old	2: Instructor
Man	White	Women-Dominated	Incongruent	Old	2: Instructor
Man	White	Neutral	Neutral	Old	2: Instructor
Man	Racially/ethnically minoritized	Man-Dominated	Congruent	Old	2: Instructor
Man	Racially/ethnically minoritized	Women-Dominated	Incongruent	Old	2: Instructor
Man	Racially/ethnically minoritized	Neutral	Neutral	Old	2: Instructor
Woman	White	Man-Dominated	Incongruent	Old	2: Instructor
Woman	White	Women-Dominated	Congruent	Old	2: Instructor
Woman	White	Neutral	Neutral	Old	2: Instructor

[17] Due to confidentiality concerns, the specific disciplines and course levels/numbers are not provided as there are some courses or even levels of courses which are only taught by one or two instructors. Therefore revealing the discipline and course level/number would reveal the identity of the instructor. This is especially a concern with respect to racially/ethnically minoritized instructors who are a very small portion of the overall population and thus easily identifiable if even the discipline is revealed. More details about the specific courses included in the analyses are available, however every attempt was made to ensure that each group of selected responses was as similar across course characteristics including discipline and course level as possible.

Woman	Racially/ethnically minoritized	Man-Dominated	Incongruent	Old	2: Instructor
Woman	Racially/ethnically minoritized	Women-Dominated	Congruent	Old	2: Instructor
Woman	Racially/ethnically minoritized	Neutral	Neutral	Old	2: Instructor
Man	White	Man-Dominated	Congruent	New	1: Helped Learn
Man	White	Women-Dominated	Incongruent	New	1: Helped Learn
Man	White	Neutral	Neutral	New	1: Helped Learn
Man	Racially/ethnically minoritized	Man-Dominated	Congruent	New	1: Helped Learn
Man	Racially/ethnically minoritized	Women-Dominated	Incongruent	New	1: Helped Learn
Man	Racially/ethnically minoritized	Neutral	Neutral	New	1: Helped Learn
Woman	White	Man-Dominated	Incongruent	New	1: Helped Learn
Woman	White	Women-Dominated	Congruent	New	1: Helped Learn
Woman	White	Neutral	Neutral	New	1: Helped Learn
Woman	Racially/ethnically minoritized	Man-Dominated	Incongruent	New	1: Helped Learn
Woman	Racially/ethnically minoritized	Women-Dominated	Congruent	New	1: Helped Learn
Woman	Racially/ethnically minoritized	Neutral	Neutral	New	1: Helped Learn
Man	White	Man-Dominated	Congruent	New	2: Change
Man	White	Women-Dominated	Incongruent	New	2: Change
Man	White	Neutral	Neutral	New	2: Change
Man	Racially/ethnically minoritized	Man-Dominated	Congruent	New	2: Change
Man	Racially/ethnically minoritized	Women-Dominated	Incongruent	New	2: Change
Man	Racially/ethnically minoritized	Neutral	Neutral	New	2: Change
Woman	White	Man-Dominated	Incongruent	New	2: Change
Woman	White	Women-Dominated	Congruent	New	2: Change
Woman	White	Neutral	Neutral	New	2: Change
Woman	Racially/ethnically minoritized	Man-Dominated	Incongruent	New	2: Change

Woman	Racially/ethnically minoritized	Women-Dominated	Congruent	New	2: Change
Woman	Racially/ethnically minoritized	Neutral	Neutral	New	2: Change

Appendix 18: Average Quantitative Code Score by Instructor Group and Question

Time Period	Instructor Group	Level of Congruity	Question	Positive Professional	Negative Professional	Positive Personal	Negative Personal	Positive Course	Negative Course	Avg Of Codes	Avg Overall Score	Avg Net Score
Old	White Man Man-Dominated	Congruent	1 Course	0.00	-0.07	0.00	0.00	0.87	-0.43	0.06	0.37	1.37
Old	White Man Man-Dominated	Congruent	2 Instructor	1.37	-0.57	0.40	-0.13	0.03	-0.23	0.14	0.87	2.87
New	White Man Man-Dominated	Congruent	1 Helped Learn	0.40	-0.03	0.07	-0.03	1.37	-0.03	0.29	1.73	1.93
New	White Man Man-Dominated	Congruent	2 Change	0.13	-0.47	0.00	-0.10	0.10	-1.93	-0.38	-2.23	2.73
Old	White Man Woman-Dominated	Incongruent	1 Course	0.00	-0.07	0.10	0.00	1.07	-0.37	0.12	0.73	1.60
Old	White Man Woman-Dominated	Incongruent	2 Instructor	1.23	-0.53	0.33	-0.27	0.10	-0.47	0.07	0.40	2.93
New	White Man Woman-Dominated	Incongruent	1 Helped Learn	0.50	0.00	0.07	0.00	1.27	-0.07	0.29	1.77	1.90
New	White Man Woman-Dominated	Incongruent	2 Change	0.00	0.00	0.00	-0.03	0.13	-0.60	-0.08	-0.50	0.77
Old	White Man Neutral	Neutral	1 Course	0.03	-0.03	0.00	0.00	0.83	-0.63	0.03	0.20	1.53
Old	White Man Neutral	Neutral	2 Instructor	1.87	-0.13	0.67	-0.03	0.00	-0.03	0.39	2.33	2.73
New	White Man Neutral	Neutral	1 Helped Learn	0.53	0.00	0.20	0.00	1.13	-0.03	0.31	1.83	1.90

New	White Man Neutral	Neutral	2 Change	0.03	-0.20	0.00	0.00	0.10	-1.23	-0.22	-1.30	1.57
Old	White Woman Man-Dominated	Incongruent	1 Course	0.03	-0.07	0.00	0.00	0.90	-0.50	0.06	0.30	1.50
Old	White Woman Man-Dominated	Incongruent	2 Instructor	1.80	-0.27	0.50	0.00	0.00	-0.10	0.32	1.93	2.67
New	White Woman Man-Dominated	Incongruent	1 Helped Learn	0.70	-0.30	0.13	0.00	0.93	-0.13	0.22	1.33	2.20
New	White Woman Man-Dominated	Incongruent	2 Change	0.03	-0.63	0.10	-0.10	0.07	-0.73	-0.21	-1.27	1.67
Old	White Woman Woman-Dominated	Congruent	1 Course	0.00	0.00	0.00	0.00	1.03	-0.97	0.01	0.07	2.00
Old	White Woman Woman-Dominated	Congruent	2 Instructor	1.43	-0.37	0.33	0.00	0.03	-0.27	0.19	1.17	2.43
New	White Woman Woman-Dominated	Congruent	1 Helped Learn	0.57	0.00	0.03	0.00	1.43	-0.17	0.31	1.83	2.20
New	White Woman Woman-Dominated	Congruent	2 Change	0.07	-0.20	0.00	0.00	0.37	-0.87	-0.11	-0.63	1.50
Old	White Woman Neutral	Neutral	1 Course	0.07	0.00	0.00	0.00	0.87	-0.43	0.08	0.50	1.37
Old	White Woman Neutral	Neutral	2 Instructor	1.40	-0.23	0.60	-0.03	0.17	-0.10	0.30	1.80	2.53
New	White Woman Neutral	Neutral	1 Helped Learn	0.24	0.00	0.03	0.00	1.21	-0.07	0.24	1.41	1.50
New	White Woman Neutral	Neutral	2 Change	0.07	-0.30	0.00	0.00	0.03	-1.10	-0.22	-1.30	1.50

Old	Racially/ethnically Minoritized Man Man-Dominated	Congruent	1 Course	0.00	-0.13	0.00	-0.07	0.67	-0.83	-0.06	-0.37	1.70
Old	Racially/ethnically Minoritized Man Man-Dominated	Congruent	2 Instructor	1.33	-0.60	0.43	0.00	0.33	-0.43	0.18	1.07	3.13
New	Racially/ethnically Minoritized Man Man-Dominated	Congruent	1 Helped Learn	0.27	0.00	0.07	0.00	1.30	-0.07	0.26	1.57	1.70
New	Racially/ethnically Minoritized Man Man-Dominated	Congruent	2 Change	0.20	-0.17	0.03	-0.03	0.10	-1.00	-0.14	-0.87	1.53
Old	Racially/ethnically Minoritized Man Woman-Dominated	Incongruent	1 Course	0.07	-0.03	0.00	0.00	0.83	-0.63	0.04	0.23	1.57
Old	Racially/ethnically Minoritized Man Woman-Dominated	Incongruent	2 Instructor	1.13	-0.93	0.30	-0.30	0.43	-0.23	0.07	0.40	3.33
New	Racially/ethnically Minoritized Man Woman-Dominated	Incongruent	1 Helped Learn	0.33	0.00	0.20	0.00	1.33	-0.03	0.31	1.83	1.90
New	Racially/ethnically Minoritized Man Woman-Dominated	Incongruent	2 Change	0.20	-0.23	0.00	0.00	0.10	-0.83	-0.13	-0.77	1.37
Old	Racially/ethnically Minoritized Man Neutral	Neutral	1 Course	0.03	-0.17	0.00	0.00	0.70	-0.73	-0.03	-0.17	1.63

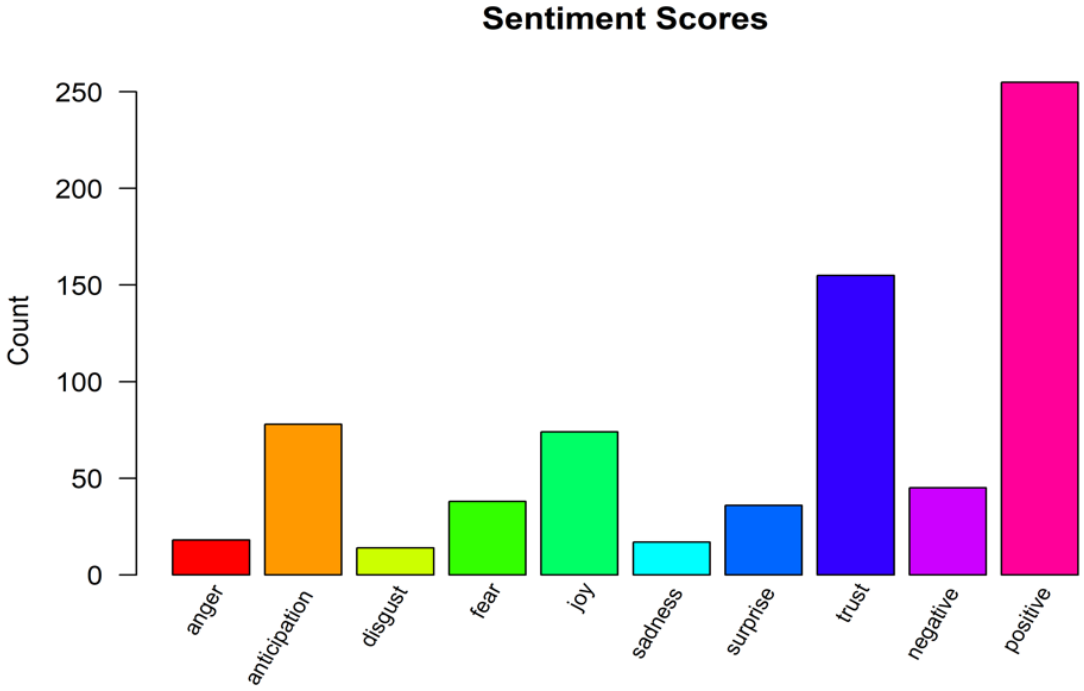
Old	Racially/ethnically Minoritized Man Neutral	Neutral	2 Instructor	1.40	-0.53	0.67	-0.17	0.23	-0.33	0.21	1.30	3.33
New	Racially/ethnically Minoritized Man Neutral	Neutral	1 Helped Learn	0.57	-0.20	0.07	0.00	0.80	-0.20	0.17	1.03	1.83
New	Racially/ethnically Minoritized Man Neutral	Neutral	2 Change	0.07	-0.43	0.00	0.00	0.07	-0.80	-0.18	-1.10	1.37
Old	Racially/ethnically Minoritized Woman Man-Dominated	Incongruent	1 Course	0.03	-0.10	0.00	0.00	0.80	-0.60	0.02	0.13	1.53
Old	Racially/ethnically Minoritized Woman Man-Dominated	Incongruent	2 Instructor	1.50	-1.30	0.40	-0.37	0.07	-0.37	-0.01	-0.07	4.00
New	Racially/ethnically Minoritized Woman Man-Dominated	Incongruent	1 Helped Learn	0.57	-0.03	0.13	0.00	1.23	-0.03	0.31	1.87	2.00
New	Racially/ethnically Minoritized Woman Man-Dominated	Incongruent	2 Change	0.07	-0.37	0.00	0.00	0.23	-0.70	-0.13	-0.77	1.37
Old	Racially/ethnically Minoritized Woman Woman-Dominated	Congruent	1 Course	0.00	0.00	0.00	0.00	0.63	-0.83	-0.03	-0.20	1.60
Old	Racially/ethnically Minoritized	Congruent	2 Instructor	1.67	-0.43	0.50	-0.27	0.20	-0.20	0.24	1.47	3.27

	Woman Woman-Dominated											
New	Racially/ethnically Minoritized Woman Woman-Dominated	Congruent	1 Helped Learn	0.53	-0.10	0.30	0.00	0.97	-0.07	0.27	1.63	1.97
New	Racially/ethnically Minoritized Woman Woman-Dominated	Congruent	2 Change	0.07	-0.10	0.00	0.00	0.07	-1.20	-0.19	-1.50	1.77
Old	Racially/ethnically Minoritized Woman Neutral	Neutral	1 Course	0.04	0.00	0.00	0.00	0.64	-0.84	-0.03	-0.16	1.52
Old	Racially/ethnically Minoritized Woman Neutral	Neutral	2 Instructor	1.20	-0.80	0.40	-0.12	0.24	-0.24	0.11	0.68	3.00
New	Racially/ethnically Minoritized Woman Neutral	Neutral	1 Helped Learn	0.80	-0.10	0.17	-0.03	0.97	-0.13	0.28	1.73	2.27
New	Racially/ethnically Minoritized Woman Neutral	Neutral	2 Change	0.07	-0.47	0.00	0.00	0.07	-0.80	-0.19	-1.13	1.40
Average Of Old Questions				0.74	-0.31	0.23	-0.07	0.49	-0.45	0.10	0.62	2.30
Average Of New Questions				0.29	-0.18	0.07	-0.01	0.64	-0.53	0.05	0.26	1.74
Average Of All Questions				0.51	-0.24	0.15	-0.04	0.56	-0.49	0.07	0.44	2.02
Number Of Categories with Score Of Zero				6.00	11.00	21.00	32.00	2.00	0.00	0.00	0.00	0.00

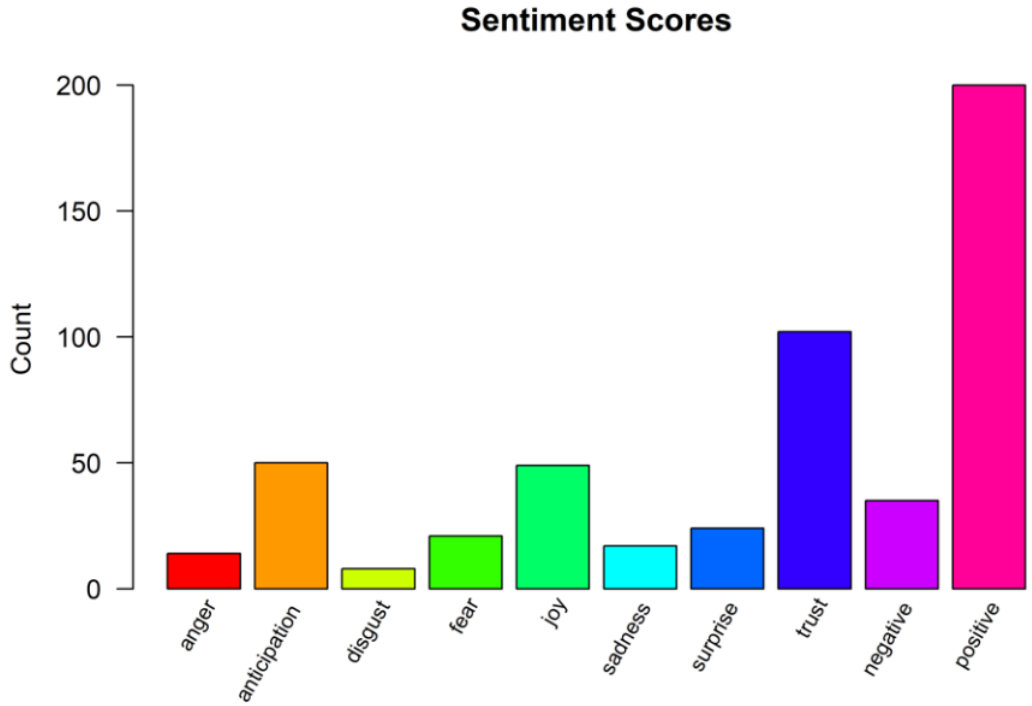
Appendix 19: Sentiment Analyses

Role Congruent Instructors

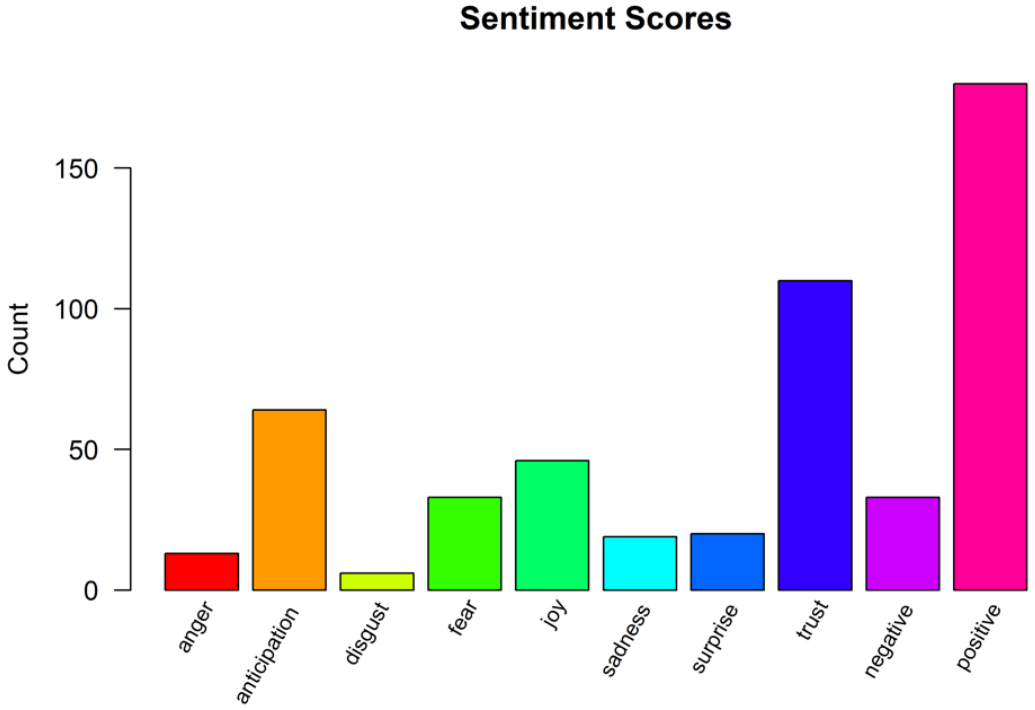
White Men Man-Dominated Discipline



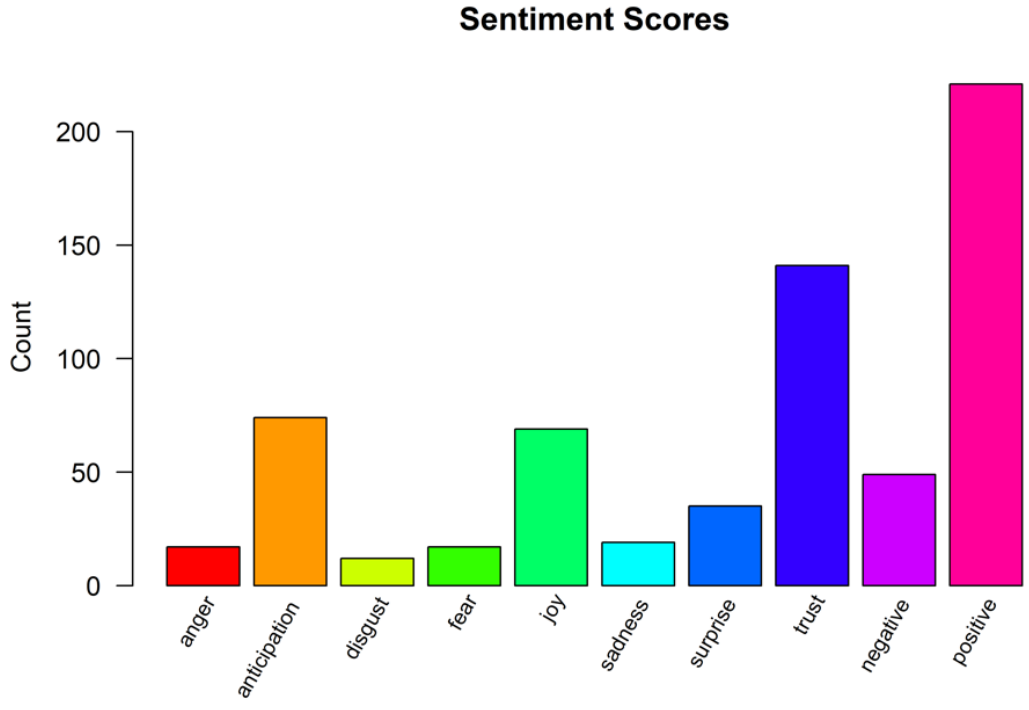
White Women Woman-Dominated Discipline



Racially/ethnically Minoritized Men Man-Dominated Discipline

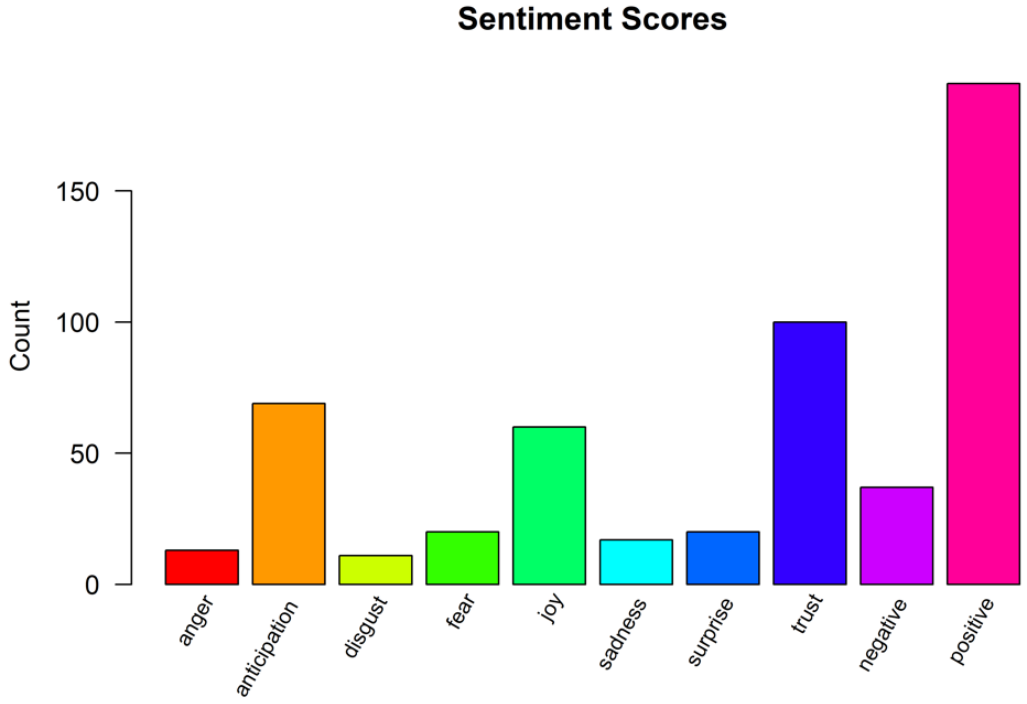


Racially/ethnically Minoritized Women Woman-Dominated Discipline

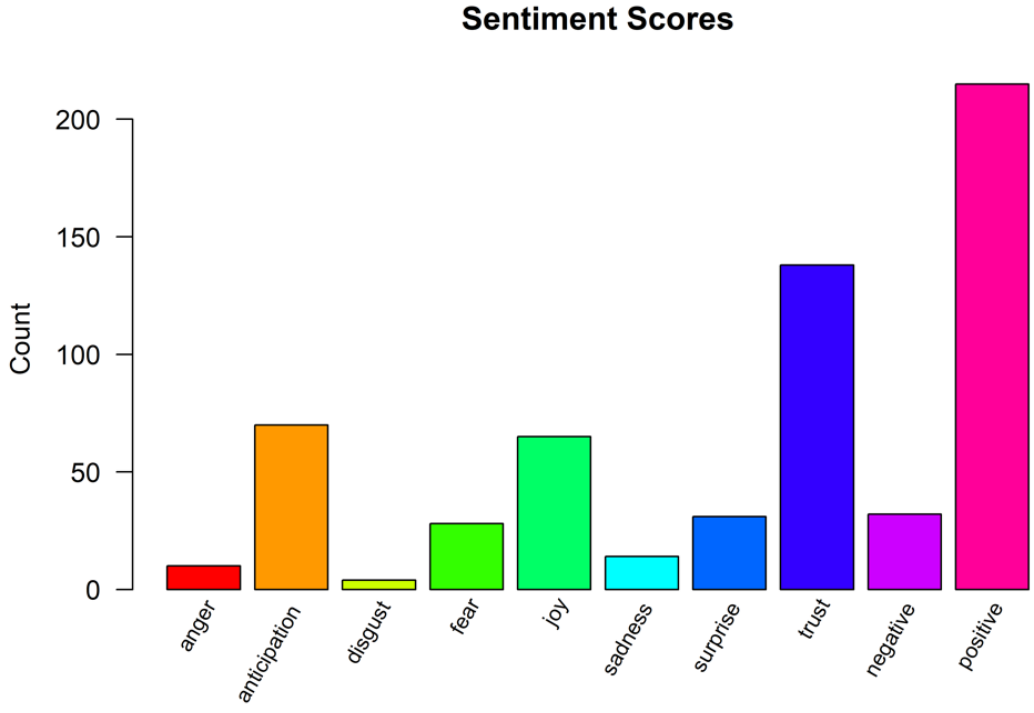


Role Incongruent Instructors

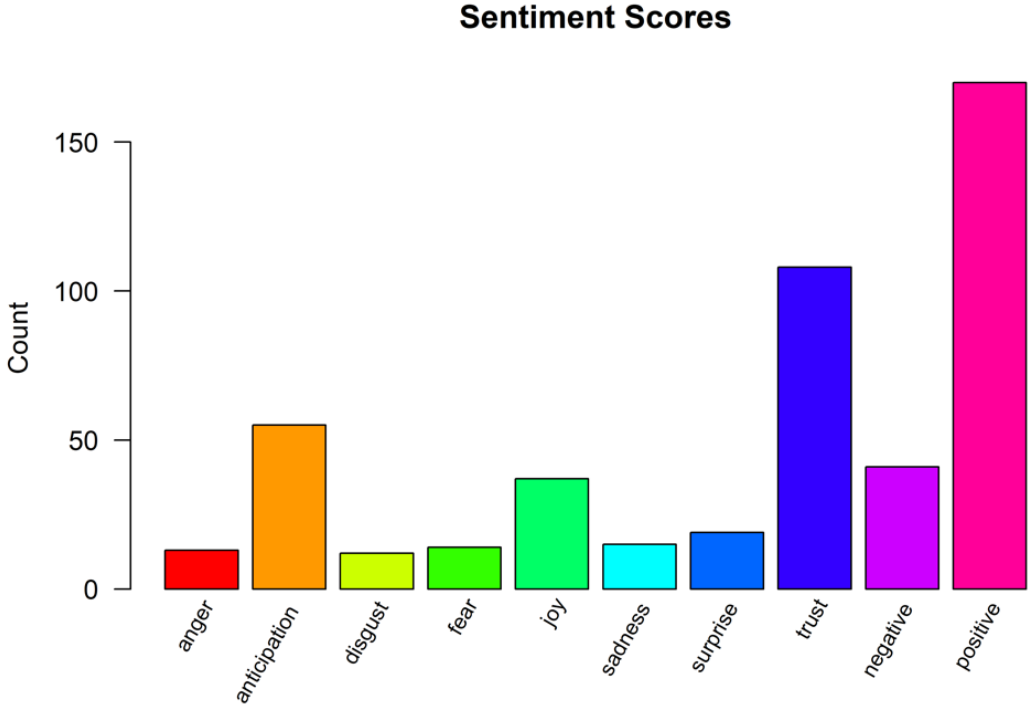
White Man Woman-Dominated Discipline



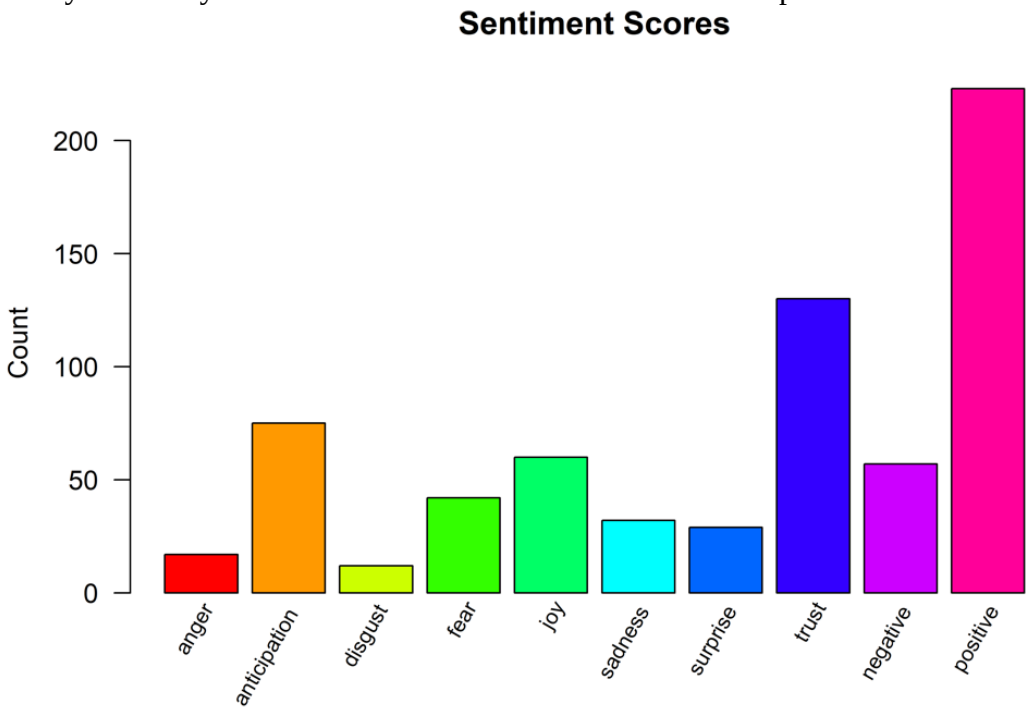
White Women Man-Dominated Discipline



Racially/ethnically Minoritized Men Woman-Dominated Discipline

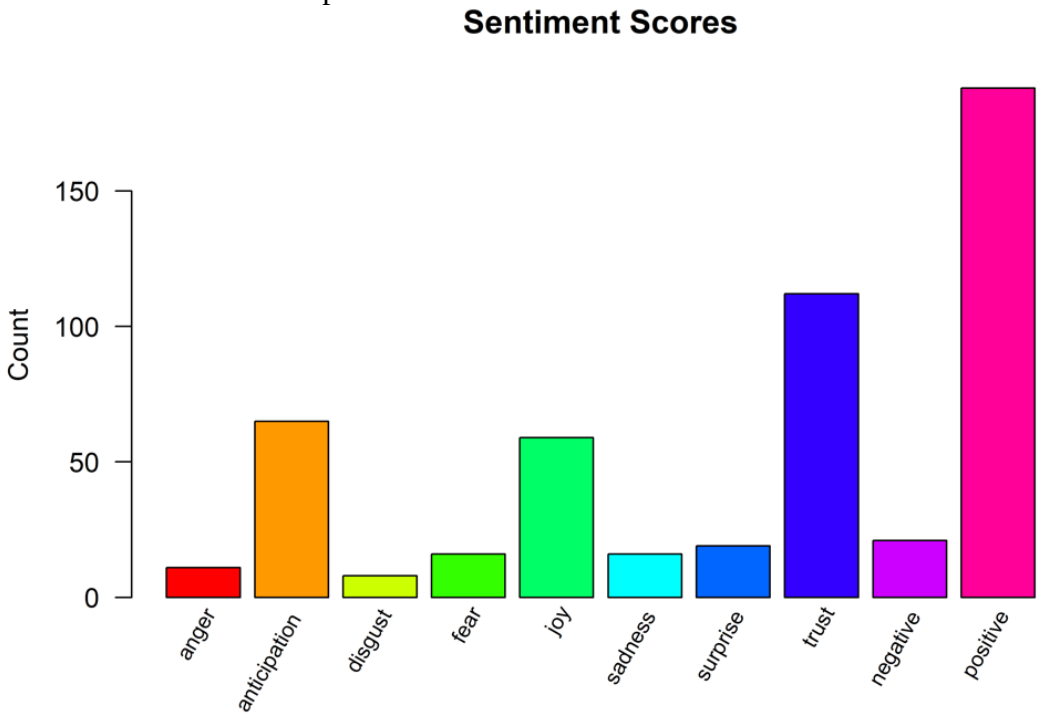


Racially/ethnically Minoritized Women Man-Dominated Discipline

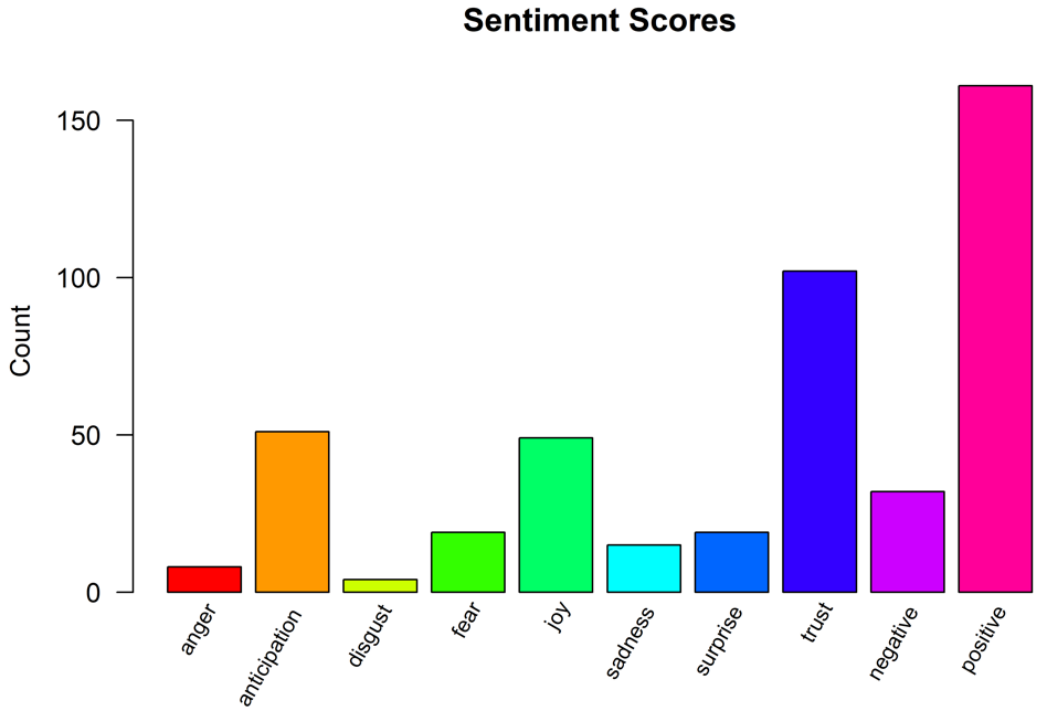


Role Neutral Instructors

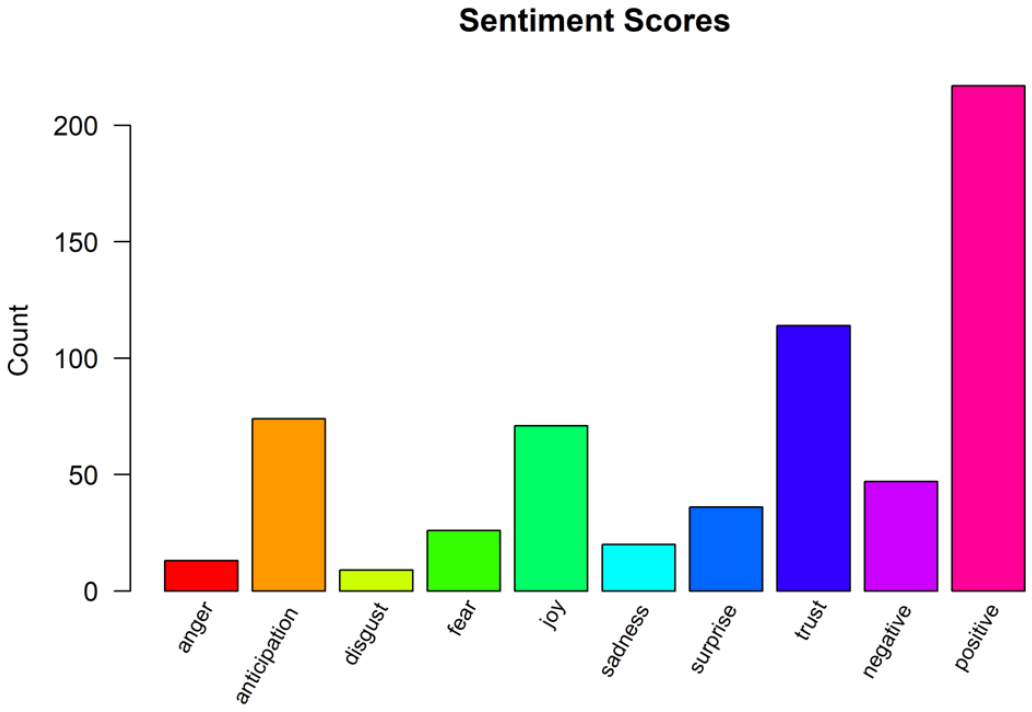
White Men Neutral Discipline



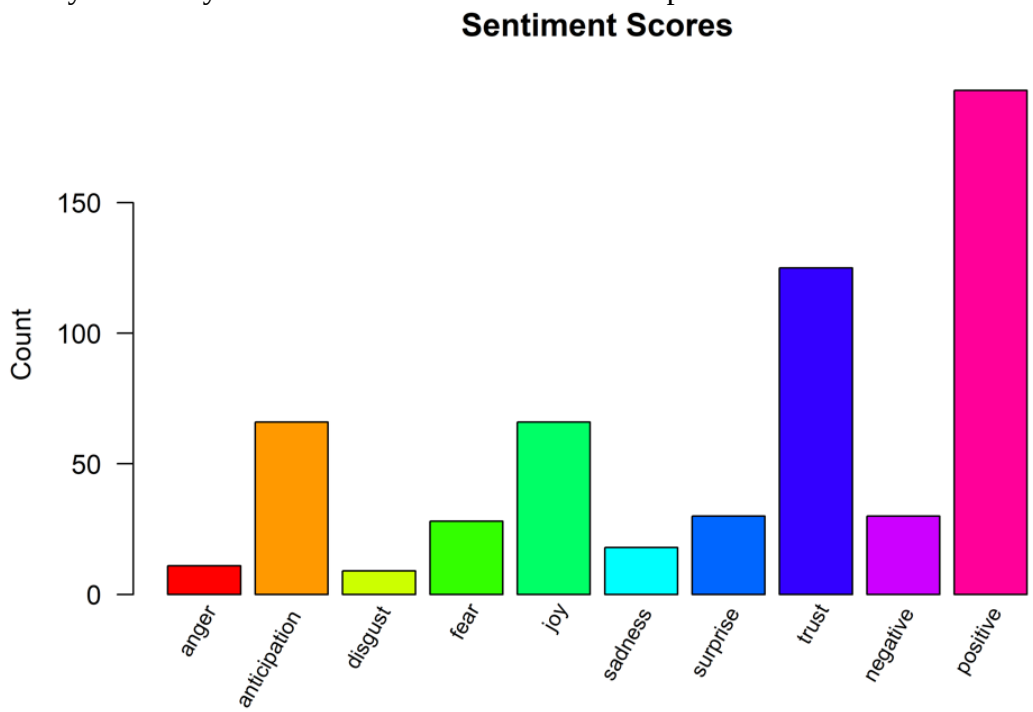
White Women Neutral Discipline



Racially/ethnically Minoritized Men Neutral Discipline



Racially/ethnically Minoritized Women Neutral Discipline

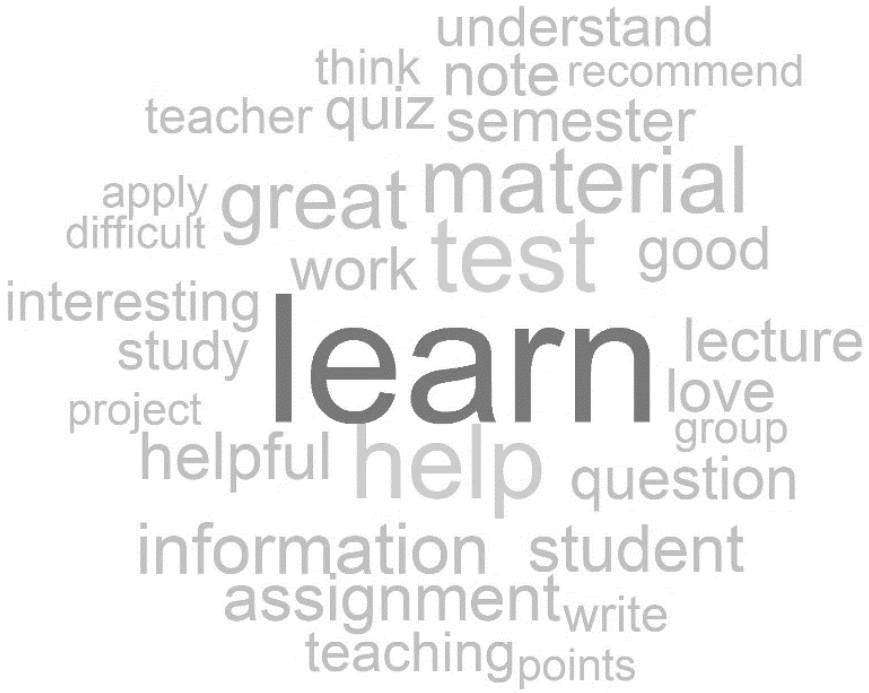


Appendix 20: Word clouds

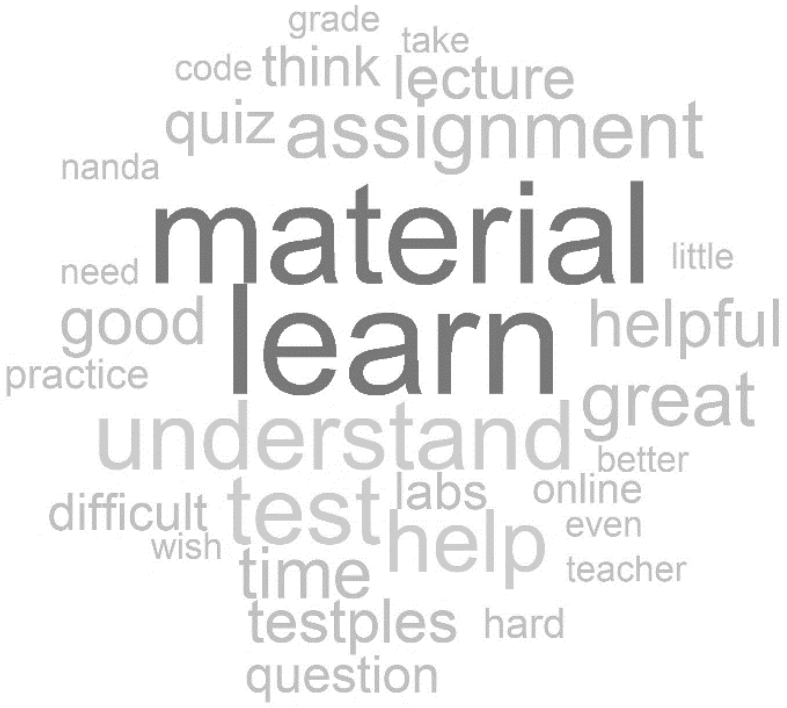
Role Congruent Instructors

White Men Man-Dominated Discipline

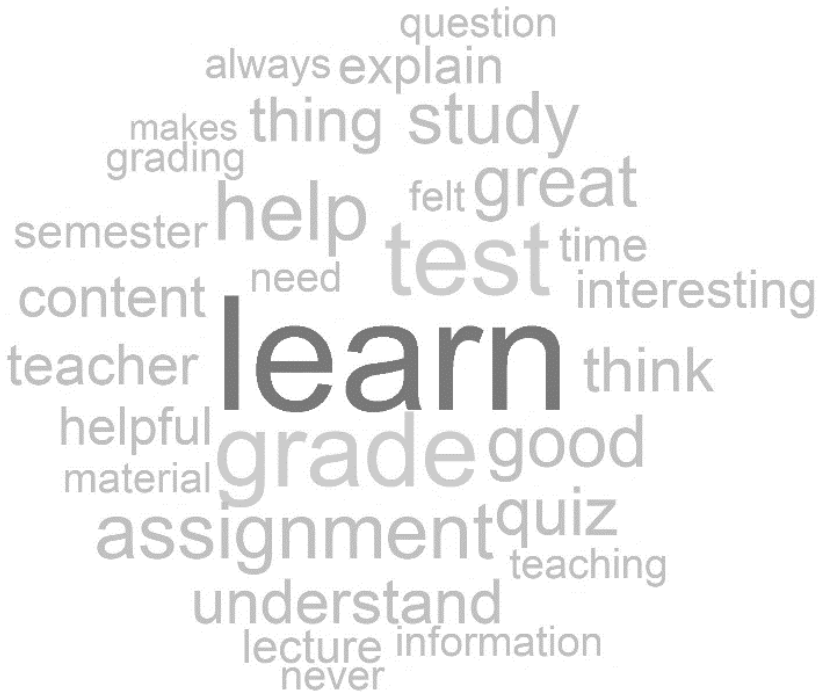
White Women Woman-Dominated Discipline



Racially/ethnically Minoritized Men Man-Dominated Discipline

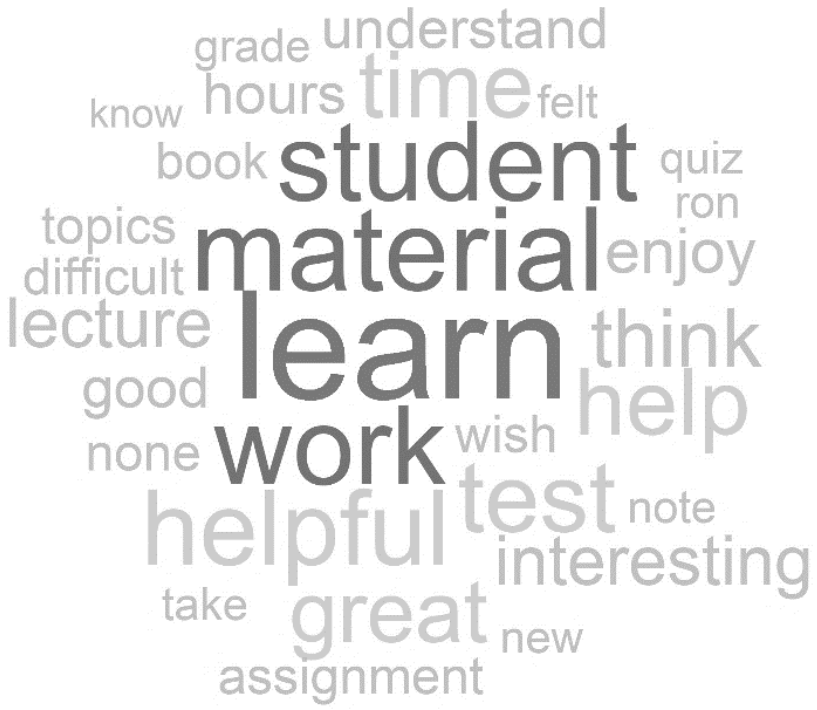


Racially/ethnically Minoritized Women Woman-Dominated Discipline

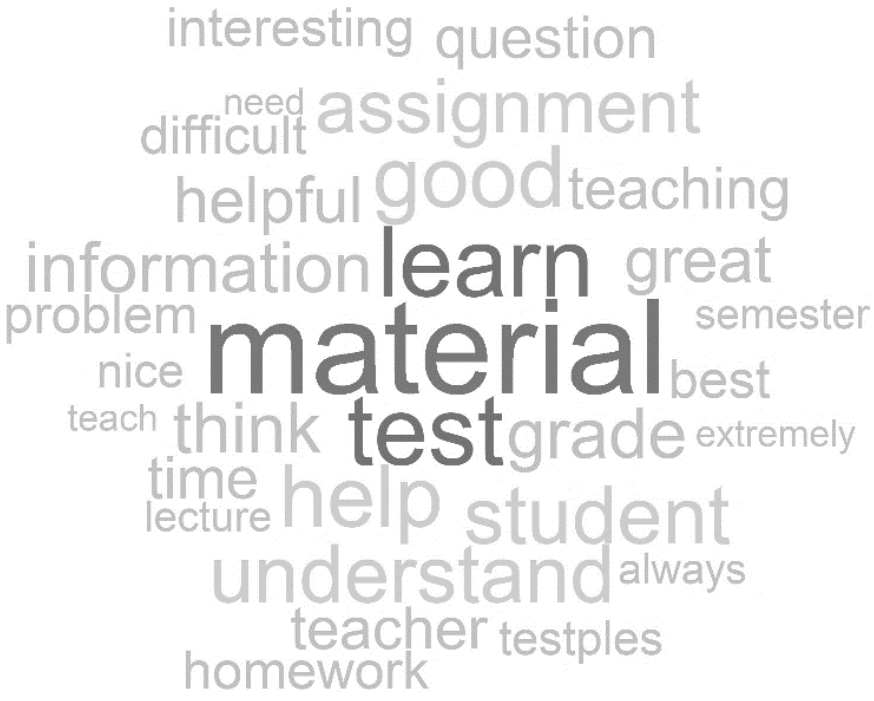


Role Incongruent Instructors

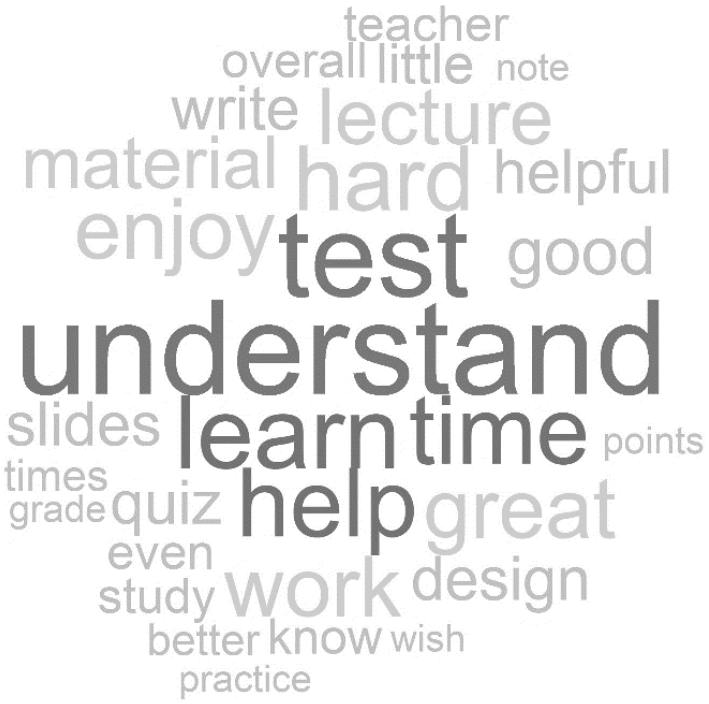
White Men Woman-Dominated Discipline



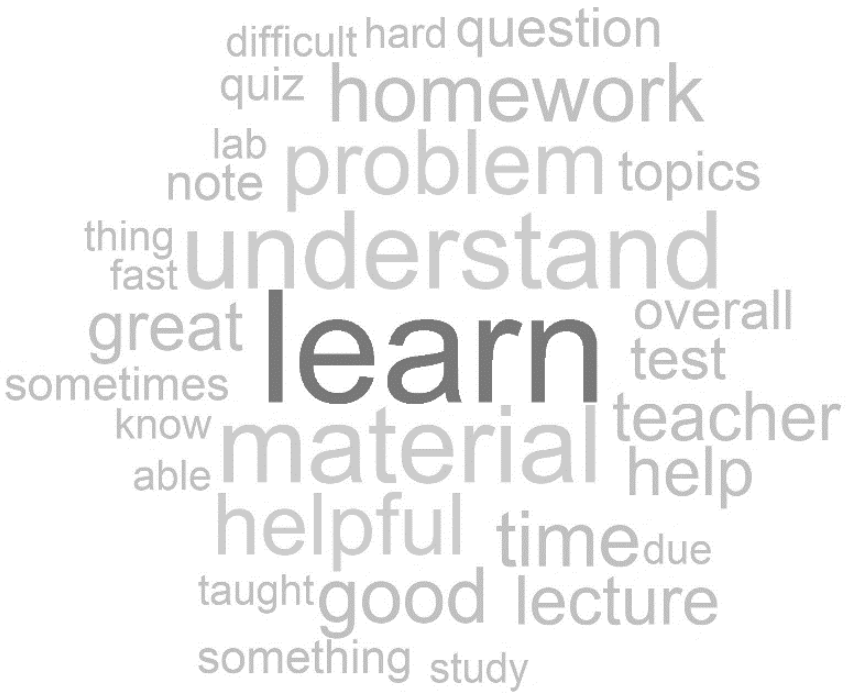
White Women Man-Dominated Discipline



Racially/ethnically Minoritized Men Man-Dominated Discipline



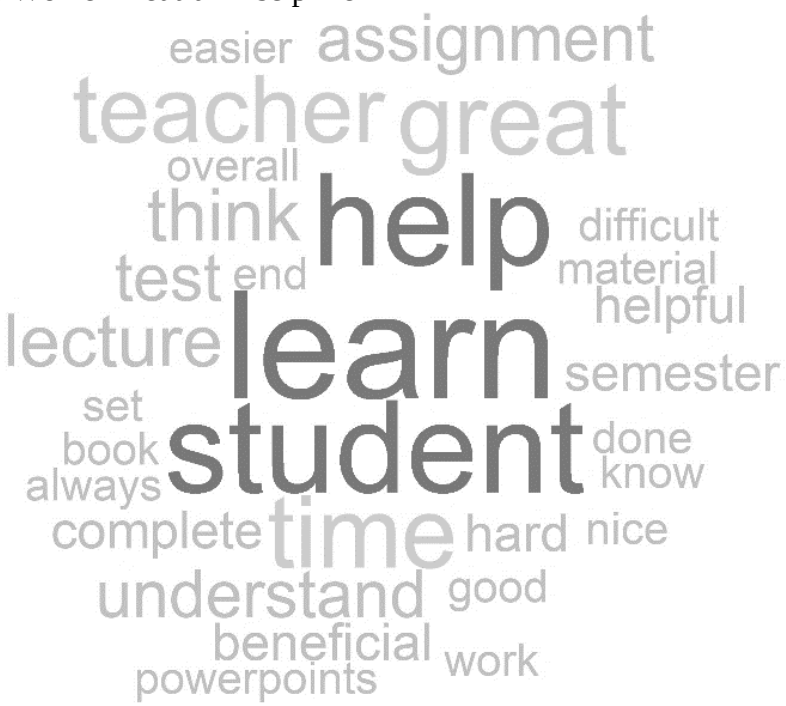
Racially/ethnically Minoritized Women Woman-Dominated Discipline



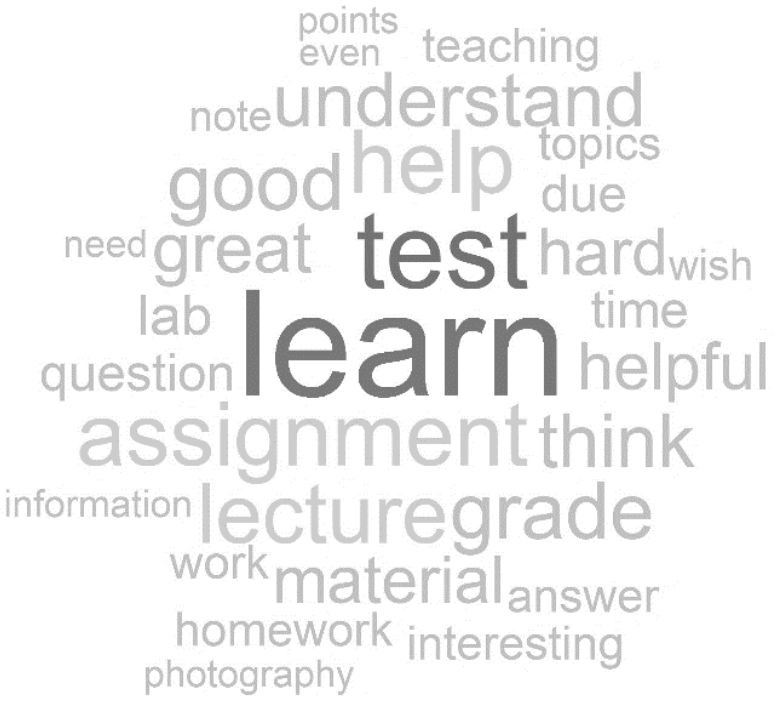
Role Neutral Instructors

White Men Neutral Discipline

White Women Neutral Discipline



Racially/ethnically Minoritized Men Neutral Discipline



Racially/ethnically Minoritized Women Neutral Discipline[18]

[18] *Note:* Black box is covering the name of an instructor whose name was in enough reviews to make it into the list of most-used words

Appendix 21: Qualitative Code Themes by Instructor Group

Code	White Man Man-Dominated Summary (Role Congruent)
Positive Professional	Intelligent, expert, knowledgeable, good at teaching, helpful, responsive, having clear expectations, approachable, well versed in the materials, and having good pace and presentations.
Negative Professional	Lack of study guide, required self-studying, unclear, not a good teacher, not entertaining, boring, fast, picked on students, disorganized, lack of notice on assignments, and bad at explaining.
Positive Personal	Helpful, nice, knowledgeable, kindhearted, caring, resourceful, and good sense of humor.
Negative Personal	Jerk, off-putting, rude, could not casually converse with them, sporadic, easily distracted, and repeated the word "ultimately".
Positive Course	Good, great, helpful, interesting, developed skills, teamwork, communication, valuable information, and good review sheets, assignments, quizzes, take-home tests, textbooks, lectures, homework, in-class assignments, and hands-on practice.
Negative Course	Not helpful, did not learn, outdated assignment/materials, needed more guidance on assignments, disorganized, bad notes, and desire for study guides, different exams, attendance points, hands-on activities, clearer grading criteria, more structure, a better syllabus, less repetition of other courses, more organizations, and examples.

Code	White Woman Woman-Dominated Summary (Role Congruent)
Positive Professional	Helpful, excellent, fair, reasonable, knowledgeable, passionate, good teaching style, well-paced, good examples, thorough, fantastic, asset to institution, good feedback, responsive, created engaging/positive learning environment, and knowledgeable.
Negative Professional	Hard to learn from, boring, unclear, a harsh grader, went off topic, not prepared, used filler words when talking, did not use technology well (including bad PowerPoints), unclear guidelines, and need to help more with studying.
Positive Personal	Calm, enthusiastic, nice, approachable, energetic, and collected.
Negative Personal	No comments.
Positive Course	Good, relevant, interesting, enjoyable materials (including lectures, readings, quizzes, examples, assignments, PowerPoints, study guides, quizzes, discussions, and class Google Drive), wonderful, great, not needing changes, and would recommend the class to others anyone could learn from it and enjoy it.
Negative Course	Too much work (including group work and note cards), did not provide study guides, bad materials (including tricky test questions), repetitive material, desire for more materials/information (more details in PowerPoints, clicker questions, quizzes, more homework, in-class activities, more details about assignments sooner, less reading, and more videos), and dislike for administrative portions of the class (mandatory, not offered every semester, attendance policy, and lack of points).

Code	Racially/ethnically Minoritized Men Man-Dominated Summary (Role Congruent)
Positive Professional	Amazing, "one of the best", passionate, knowledgeable, available, helpful, pushes/wants students to learn, good, went over material slowly, good examples, phenomenal, applied materials, and should not change a thing.
Negative Professional	Not effective, bad teaching style, lectures were redundant of textbook, grades were not posted timely, uncommunicative, gave non-straightforward answers, monotone, hard to follow/understand, and unavailable outside of class.
Positive Personal	Great guy, humorous, charismatic, cool, easy to talk to, helpful, used personal time to help students, and enjoyable person.
Negative Personal	Language barrier and did not care about students or their learning.
Positive Course	Great, easy, interesting, fun, beneficial, good pace, well designed, enjoyable, would not change a thing, good materials (practice problems, videos, hands-on in-class activities, homework, slides, examples, discussions, and quizzes), and had a good grading system.
Negative Course	Disorganized, poorly designed, useless, not helpful, too much work, covered too much information, hard, not enough examples/practice problems, slides were bad and/or not made available to students, students felt they had to teach themselves, bad TA, needed more assignments and/or assignments to be introduced sooner, dry material, too much emphasis on memorizing, and assignments were emailed instead of posted to the learning management system.

Code	Racially/ethnically Minoritized Women Women-Dominated Summary (Role Congruent)
Positive Professional	Great, knowledgeable, fair, skilled, cared if they learned, willing to help, knew the course material, gave good explanations, taught at a good pace, made the class interesting, learning environment was very welcoming, organized, excellent, communicative, positive, passionate about course material, engaged, amazing, supportive, and "one of the best", pushed students to do their best, and always made sure everyone was doing well.
Negative Professional	Made the course harder than it needed to be, did not always know what was going on, unfair graders, taught only their perspective, unprofessional, argumentative when presented with opposing views, disorganized, everything was incorrect, would not recommend, learning environment was unwelcoming, test questions were tricky, should not require the purchase of online access codes, should offer retakes of quizzes, went too fast, bad explanations, did not ask for differing opinions, did not utilize the learning management system, did not post grades online, explanations for upcoming projects were lacking, and "not need to be a professor".

Positive Personal	Enthusiastic, passionate, friendly, awesome, nice, helpful, caring, encouraging, energetic, sweet, funny, helpful, understanding, approachable, compassionate, positive, turned their students into a better person, and did not make students feel stupid.
Negative Personal	Three no comments, Instructor comments noted they were unprofessional, rude, mean, and biased towards their opinions.
Positive Course	Good, interesting, eye-opening, fun, worth learning, enjoyed the materials, learned a lot, good materials (television shows/videos, PowerPoints/lectures, review sessions, quizzes, PowerPoints, extra credit, discussions, readings, self-assessments, clicker questions, and writing exercises), organized, great, and everything helped with learning.
Negative Course	Not well planned, too many topics, waste of time and money, not necessary, assignments were not always clear, hard, unpredictable, about topics only mentioned briefly, bad lectures/PowerPoints, nothing helped with learning, needed more materials (clear study guides, exams, clicker questions, less group work, required book) unfair grading, bad explanations of assignments, not fun, did not want to pay as much for online access, material was not well calibrated for the students in the course, and "awful just awful".

Code	**White Man Woman-Dominated Summary (Role Incongruent)**
Positive Professional	Good, fair grader, helpful, clear, knowledgeable, flexible, good teaching style, dedicated, passionate, engaging, gave good examples, connected material to real-life applications, gave good feedback, and created a good learning environment.
Negative Professional	Too fast, did not provide slides, did not learn from them, bad teaching style, made the course difficult, did not respond to emails, and did not provide timely or useful feedback.
Positive Personal	Made the course worth it, nice, passionate, good to talk to, helpful, funny, cares about their students, and positive and relaxed but strict in a good way.
Negative Personal	Rude, a pushover, intense, and lacked care and respect for their students.
Positive Course	Fun, interesting, worthwhile, good content/materials (including textbook, visuals, examples, quizzes, discussions, readings, lectures, agendas, study guides, projects, and out of class work days), good assignments, awesome, learned a lot, and students would not make changes.
Negative Course	Hard, difficult, required a lot of time outside of class, disappointed in topics and texts, not enough points, not enough time for questions/help, students felt they learned more from other students/online, and desire for more quizzes, assignments, examples, and grades/assignments to be available online.

Code	White Women Man-Dominated Summary (Role Incongruent)
Positive Professional	Created comfortable learning environment, great, willing to answer questions, responsive to emails, clear, understanding, knowledgeable, fair, helpful, wanted students to learn, challenged students to think, provided useful materials, passionate, "one of the best", accommodating, organized, good teaching style (stories, examples, explanations), made course intriguing and interactive, incorporated jokes well, and "keep up the good work".
Negative Professional	Bad teaching style, assignments lacked detail, quiet, one instructor went too slow and did not explain things well while another went too fast and over-explained, taught themselves, "worst professor", provided irrelevant information, did not grade timely, "past the point of being an effective professor", lacked understanding, terrible, graded harshly, only gave negative feedback, missed a lot of class, and too smart for their own good.
Positive Personal	Caring, role model, approachable, nice, "nice lady", knowledgeable, helpful, understating, awesome, and compassionate.
Negative Personal	Three questions had no comments, two comments on Change—condescending, rude, and nice in class but different when meeting about assignments.
Positive Course	Interesting, good materials (including textbook, discussions, homework assignments, in-class examples, demonstrations, videos, group project, PowerPoints, quizzes, lectures, and review sessions), good hands-on experiences, enjoyable, useful, well-organized, enjoyable, and nothing should change.
Negative Course	Bad materials (including the question types on tests, clicker questions, theory focus, generalized lectures, dry material, and long assignments), some felt too much work while others wanted more, difficult, jumbled topics, lectures and tests did not align, internet/self-study helped with learning, lack of examples, too much writing, lack of clarity on textbook edition, and desire for notes to be shared.

Code	Racially/ethnically Minoritized Men Woman-Dominated Summary (Role Incongruent)
Positive Professional	Great, excellent, good, knowledgeable, amazing, helpful, talented, gave good feedback, provided good explanations, gave good advice, made course enjoyable, helped students learn, caring, wonderful, fantastic, enthusiastic, intelligent, willing to go the extra mile, responsive, and taught well.
Negative Professional	Did not take lack of prior knowledge into consideration, course pace was too fast, did not give enough feedback, grades harshly, unclear, thought they knew more than everyone, hard to talk to, hard to follow, bad explanations, not engaging, dry, too high of expectations, and tough to the point of unprofessional.

Positive Personal	Great guy, bright, upbeat, laid back, easy to communicate with, amazing, positive, encouraging, understanding, helpful, and used personal time to help them.
Negative Personal	Only comments on Instructor question. Dry, rude, hard to understand, stubborn, arrogant, and hard to get along with.
Positive Course	Good, useful, interesting, enjoyable, helped students decide on their career paths, helped establish better working habits, worthwhile challenge, learned a lot from it, good materials (guest lectures, lab reports, practice tests, test review, studio sessions, PowerPoints, case studies, discussions, textbook, examples, and in-class activities), good structure, fair tests, and wish there were more classes like this.
Negative Course	Bad materials/assignments, too much work required outside of class, useless, unenjoyable, "my own personal hell", nothing helped learning, hard, bad grading structure, would not recommend, hard tests, bad PowerPoints, needed more content (quizzes, review sessions, longer class sessions, examples, and activities), course moved too quickly, difficult to know the depth at which they needed to learn the materials, and wish they had not taken the course.

Code	**Racially/ethnically Minoritized Women Men-Dominated Summary (Role Incongruent)**
Positive Professional	Made a tough class simple, did their best, great, the best, helpful, nice, enthusiastic, caring, passionate, good at communicating, taught well, presented the information in interesting ways, genuinely interested in students' learning, liked the example problems and notes, clear, good at explaining the subject, provided helpful feedback, helpful when students asked questions, involved students in the lecture, and prompt with email responses.
Negative Professional	Did not teach well, could be better, did not give satisfactory explanations, not fair grader, did not show their notes, notes were incorrect, disorganized, not specific, punished students by curving quiz grades, repeated themselves in lectures, took too long to get into the subject., unresponsive to emails, ineffective teacher, disorganized, the "worst teacher I have ever taken", did not use class time effectively, gave assignments too close to the deadline. One instructor laughed at students' questions, threatened to have security remove students for using technology, and stared at students who left to use the restroom.
Positive Personal	Nice, caring, pleasure to be around, funny, helpful, approachable, intelligent, motivated, organized, patient, and understanding.
Negative Personal	Three no comments, comments on Instructor noted they were hard to understand (language barrier, talked too fast, and not coherent), and had bad handwriting.
Positive Course	Interesting, great, good, very important, loved the course, learned a lot, enjoyed the materials, connected well to other courses, enjoyable, good materials (activities, readings, discussions, papers, lab

	assignments, lectures, notes, videos, explanations, homework, quizzes), attending class was worth it, and nothing needed to change.
Negative Course	Unorganized, difficult, a lot of material to cover, desire for more hands-on work, lack of support for students who were struggling, needed to rely on tutors to learn, assignments were not posted timely, the section did not cover as much as others, disliked technology bans, homework questions were often ahead of the lectures, desire for more materials (homework assignments, hands-on activities), did not like quizzes at the beginning of class, homework was on multiple platforms, lecture presentation was bad, content was too complex, too fast paced, exams were difficult, and not all lecture material was relevant.

Code	White Men Neutral Summary (Role Neutral)
Positive Professional	Fine, good, great, "one of the best", available, passionate, knowledgeable, cared about students learning, well-paced, appropriate level, helped with difficult topics, created a positive learning environment, presented material well, good feedback, good teaching style, and wanted students to be present in class.
Negative Professional	Not useful, disorganized, lacked clarity, absent minded, teaching assistant did more of the teaching, strict, too fast, graded hard, and did not teach.
Positive Personal	Caring, great, funny, friendly, engaging, amazing, patient, understanding, approachable, passionate, real, straight-up, honest, and had a good sense of humor.
Negative Personal	Lacked understanding for personal circumstances.
Positive Course	Enjoyed/loved course, fine, great, good workload, materials were useful/necessary/interesting, difficult but worthwhile, and students liked various aspects of the course (including guest speakers, assignments, review sessions, homework, class discussions, labs, videos, fellow classmates, readings, and the assignment calendar).
Negative Course	Redundant of other courses, bad textbook, bad assignments such as paying to attend events and bad tests, content/assignments outdated, nothing helped students learn, not enough help sessions or instructor interaction, modality complaints (should be online, cancelled less, not full-term), should not be required, and should be more credits.

Code	White Women Neutral Summary (Role Neutral)
Positive Professional	Flexible on due dates, "one of the best teachers at the university", great, knowledgeable, engaging, helpful, helped learn/understand materials, valued their opinions, cared about the subject they were teaching, amazing, responded well to questions, good feedback, interactive teaching style, and should keep doing what they are doing.
Negative Professional	Difficult, harsh grader, high expectations, not open to differing opinions, not responsive to emails, too fast of pace, did not use

	PowerPoint, lacking explanations, not present in class, poor, made mistakes, did not teach well, and recommend staff change.
Positive Personal	Inspirational, kind, caring, lovely, beautiful, dope, fun to work with, understanding, passionate, organized, always available, and "not just a teacher but someone I can always go to".
Negative Personal	No comments on three questions, comment on Instruction "came across as bitchy".
Positive Course	Good, fine, helpful, interesting, learned a lot, necessary content, good materials (including example problems, Jeopardy, practice tests, study guides, examples, reviews, classwork, discussions, handouts, PowerPoints, quizzes, readings), and well organized.
Negative Course	Too much work (group work, long tests, long assignments), the lecture and lab did not align, bad materials (such as textbook) and many were out-of-date, not enough set due dates, unorganized, the worst, instructor expected too much prior knowledge, and students did not learn from the course.

Code	**Racially/ethnically Minoritized Men Neutral Summary (Role Neutral)**
Positive Professional	Good, great, amazing, wonderful, knowledgeable, organized, fair grader, helped students understand the material, willing to answer questions, helped students prepare for tests, good teaching/presentation style, professional, easy to approach, engaging, went above and beyond, and had a drive to see students succeed.
Negative Professional	Bad at explanations, talked in circles, bad grading procedures (unfair, too slow), poor communication (grades, due dates, general), hard to know what they wanted, picky about completion of assignments, bad teacher, hard to learn from, berated students for reaching out, "extremely unprofessional", rambled, lost the interest of their students, not the best, learned more from their peers than the instructor, did not come to class, did not give feedback, bad examples (including discriminatory examples), picked favorites, disorganized, reflected on past too much, and hard to follow.
Positive Personal	Two had no comments, one had only one comment. Friendly, smart, enthusiastic, clear, interesting, organized, honest, understanding, kind, easily reachable, helpful, charming, and witty.
Negative Personal	Three had no comments. On Instructor, hard to understand due to language barriers.
Positive Course	Good, great, interesting, useful, helpful, setup well, knew what to expect, exams were fair, lectures were clear, lectures did not rely on the PowerPoint, learned a lot, and good materials (practice exams, homework assignments, writing assignments, notes, PowerPoints, discussions, labs, course schedule, emails, and readings).
Negative Course	Work/tests were very difficult, hated the course, boring content, too fast, too much content, lack of hands-on practice, hard, disorganized, not enough opportunities to earn points, questions on quizzes/tests

	were not taught, did not need to go to class to learn the material, had to teach themselves, lectures were not helpful, too early in the morning, the material reviewed other courses, desire for more one-on-one work, need equations for exams, not enough time for projects, need more thoughtful due dates, TAs need to grade if helping in class, grades were not posted online, lectures were not interactive, materials were out of date, and materials were not available online.

Code	**Racially/ethnically Minoritized Women Neutral Summary (Role Neutral)**
Positive Professional	Taught well/good teaching style, available, offered help, concerned about student learning, great, passionate about their work, effective communicator, knowledgeable, thought-provoking, enthusiastic about the subject, inspired students to learn more and push themselves, excellent, knowledgeable, thought-provoking, good explanations, handled questions well, created a good learning environment, and fair.
Negative Professional	Did not communicate expectations well, vague, graded harshly, played favorites, did not explain concepts well, had higher expectations for their knowledge base coming into the course than what they did, taught too quickly, not helpful, not good teacher, confusing, lack of interest, did not interact much, unclear, did not go into enough depth, disorganized, did not update grades, biased, did not send out due date reminders, got out of sync with the syllabus, and frequently changed due dates from the syllabus.
Positive Personal	Nice, awesome, a lovely person, sweet, helpful, personable, kind, enthusiastic, passionate, vibrant, positive, “the bomb”, and awesome personality.
Negative Personal	Two no comments, two with one comment each. Instructor code noted they were intimidating and put down others, Helped Learn noted disorganized.
Positive Course	Useful, worth learning, not too difficult, interesting, set up well, good grading procedures, good course content (PowerPoints, practice tests, assigned projects, videos, group work, discussions, homework, in-class examples, notes, readings, quizzes, textbooks, and study guides), great, good, learned a lot, and attending class helped learning.
Negative Course	Terrible, detest, the worst, waste of time, did not like group projects, assignments were not useful, moved too quickly, not engaging, too much work assigned, need example assignments, “no”, wish someone else taught the course, nothing from the course helped learning (had to rely on self and/or friends), unstructured, disorganized, lacking content (assignments, online homework, lab practice, student involvement), wanted access to PowerPoints before class, desire for mandatory attendance, too much content on exams, and the department of the course took itself too seriously.