Development and evaluation of training to increase student perceptions of fairness in peer assessment

by

Jacklin H. Stonewall

A dissertation submitted to the graduate faculty in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Majors: Industrial Engineering and Human-Computer Interaction

Program of Study Committee:
Michael Dorneich, Major Professor
Stephen Gilbert
Michael Helwig
Sunghyun Kang
Linda Shenk

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The use of teams and team-centric pedagogies such as Team Based Learning (TBL) in classrooms has been shown to increase engagement and lead to better overall learning outcomes. Because of these positive outcomes, the use of teams is recommended in many educational fields, including engineering. For many instructors, especially those using teams, peer assessments are integral to the classroom environment as tools for both monitoring team performance and ensuring accountability. However, concerns have developed regarding the fairness of peer assessments due to student biases. In the literature, biased peer assessments have been found due to gender, race, language, peer group affiliation, socioeconomic status, and social style. However, there have not been studies that examine this issue from multiple perspectives (e.g. student, instructor, peer assessment scores) and across a wide range of academic disciplines. This work reports on such an examination as well as the development and evaluation of training to increase peer assessment fairness.

In student and instructor surveys as well as an analysis of over 20,000 peer assessment ratings across multiple academic departments, evidence of bias was found. Students and instructors both perceived bias in their classrooms and peer assessments, commonly due to gender, race, age, language, and personality. Peer assessment data itself also indicated biases due to gender, language, international student status, and race, which were largely unexplained by differences in achievement (GPA). To address these biases, peer assessment fairness training was developed. This training was initially developed using the literature and results of previous studies. A formative classroom evaluation of the training showed that while trained students were more confident in their ability to rate fairly, perceptions of fairness were unaffected. To refine the training, further requirements for its design were gathered through focus groups. These

requirements were implemented and the training underwent a summative classroom evaluation. The results of this evaluation indicated that students had higher perceptions of fairness in their peer assessments after receiving training. Students were also more confident in their and their peers' fair rating skills after receiving training. These results indicate that the training could be used broadly in classrooms to increase peer assessment fairness.

CHAPTER 1: INTRODUCTION

Objective

The objective of this work is to develop training to improve the fairness of student peer assessments of teamwork and performance in university classroom settings. This work will study the prevalence of bias in peer assessment and develop a training to mitigate these biases. Quantitative and qualitative methods were used to understand bias from the student and instructor perspectives, as well as from peer assessment data itself. This knowledge combined with information from online learning experts and instructors, the areas of psychology, sociology, and instructional and training design, informed the design of training materials to be used in classrooms that employ team work between students and ask student to assess their peers.

Active Learning Strategies in Engineering

The employment of small group, active learning strategies (such as cooperative learning or Team Based Learning) in classroom environments has been shown to increase student achievement, attendance, engagement, and lead to better overall learning outcomes (Michaelson, Knight, & Fing, 2004; Michaelson & Sweet, 2011; Allen, Copeland, Franks, Karimi, McCollum, Riese, & Lin, 2013). Because of these outcomes, team-based pedagogies and cooperative learning practices have been incorporated on college campuses as a strategy to improve the classroom engagement of underrepresented students. Indeed, research shows that learning in teams positively affects objective outcomes (such as exam scores) for minority students (Slavin & Oickle, 1981; Springer, Stanne, & Donovan, 1999).

Active Learning practices have been included in engineering education programs following recommendations from engineering professional associations like the European

Society for Engineering Education (SEFI) and the Active Learning in Engineering Education (ALE) network, and accreditation organizations such as Accreditation Board for Engineering and Technology (ABET). Enhanced learning outcomes have been demonstrated in active learning environments (Lima, Hammar Andersson & Saalman, 2016). In the particular case of engineering education, active learning has demonstrated enhanced cognitive acquisition of material over conventional lecturing approaches (Freeman, Eddy, McDonough, Smith, Okoroafor, Jordt, & Wenderoth, 2014), and has shown t disproportionate benefits for students from underrepresented minorities (Beneroso and Erans, 2020).

Peer Assessment

In many group and active learning classrooms, peer assessments are integral to ensuring individual accountability. In group learning, the use of graded assessments of peers' team contributions reduce social loafing" (failing to participate), thereby increasing individual accountability (Cestone, Levine, and Lane, 2008). Peer assessment refers to the process wherein students take part in evaluating the quality of their colleagues' learning outcomes (Sadler & Good, 2006; Topping, 2013). It is usually conducted in anonymity and supported by a grading rubric and a set of detailed instructions (Barak & Rafaeli, 2004; Dawson, 2017).

One of the most widely used team-centric teaching pedagogies is team-based learning (TBL). The prevalence of TBL in the United States has been steadily rising, especially in the medical field (Allen et al., 2013). In TBL classrooms, permanent teams are formed to maximize heterogeneity. TBL functions on four essential principles: 1) Properly formed teams that remain together for the duration of the term, 2) Readiness assurance assessments of individual's and team's pre-class preparation, 3) In class, team application exercises that promote learning of material and team development, and 4) Peer assessments designed to monitor team performance,

hold individuals accountable for their effort and contribution, and serve to improve overall team functioning (Michaelsen & Sweet, 2011).

Benefits of Peer Assessment

A growing body of research demonstrates the value of peer assessment for the learning process and the benefits of its implementation (Falchikov & Goldfinch, 2000). Peer assessment is a prominent instructional approach for increasing students' motivation (Hanrahan & Isaacs, 2001; Miedijensky & Tal, 2009; Deeley & Bovill, 2017), promoting the learning process (Barak & Rafaeli, 2004; Li & Gao, 2016) and increasing students' engagement (Bloxham & West, 2004). Taking the role of the assessor involves critical thinking (Harland, Wald, & Randhawa, 2017), and the implementation of higher order cognitive skills, such as argumentation and reasoning (Barak & Watted, 2017; Topping, 2013). Such skills might help deepen students' understanding of the scientific topic and provide them with the opportunity to reflect on their own work and improve it (Harland, Wald, & Randhawa, 2017; López-Pastor & Sicilia-Camacho, 2017; Snowball & Mostert, 2013).

These assessments have also been credited with empowering learners to engage more fully in the class (Stefani, 1994) and increasing interactions among students and between students and instructors (Wen & Tsai, 2006). Assessments can foster the development of autonomy and maturity, as well as improve social and professional skills (Topping, 1998). The process also encourages self-reflection and deeper understanding of the material, which may lead to improved retention and confidence (Langan, Wheater, Shaw, Haines, Cullen, Boyle, Penney, Oldekop, Ashcroft, & Lockey, 2005).

Communication and team cooperation skills are considered an integral part of an engineering curriculum. Trevelyan (2014) outlines this importance by highlighting that technical

collaboration occupies at least 60% of the work of engineers. Further, studies have already explored how peer assessment can be used to improve the oral competency of final-year engineering students (Kim, 2014; Liow, 2008). Learning environments employing project-based learning and peer assessment have been found to enhance engineering students' communication skills and critical thinking by enabling them to form original opinions and express individual viewpoints (Barak, Watted, & Haick, 2016; Hadim & Esche, 2002).

Fairness in Peer Assessments

Given the increasing prevalence of small group learning and a growing understanding of the benefits of peer assessments, these evaluations have become a focus of considerable amounts of research including examinations of student perceptions (Planas Lladó, Soley, Fraguell Sansbelló, Pujolras, Planella, Roura-Pascual, & Moreno; 2014) and implementation strategies (La Greca, Lai, Chan, & Herge, 2013).

However, both students and instructors have expressed concerns about the fairness of teams and their associated peer assessments, especially due to bias (Magin & Helmore, 2001; Samuel, 2004; Dancer & Kamvounias, 2005; Aryadoust, 2016). Research has shown that the experiences of women and students of color in these classrooms differ from those of their peers in terms of assessment (Wayland et al., 2014). Additionally, it is already understood that biased behaviors are commonly present in higher education classrooms (Boysen & Vogel, 2009). Because of these concerns, interest in creating fairer peer assessments has increased. The challenge in addressing these biases and creating fairer peer assessments is two-sided: the difficulty of accurately detecting and “untangling” various forms of bias in peer assessment, and the often implicit root of biased behaviors.

Bias in the Classroom

Biases, both implicit and explicit, negatively impact the way people perceive members of disadvantaged groups. Unfortunately, biases often extend into the classroom environment (Marcus, Mullins, Brackett, Tang, Allen & Pruett, 2003; Boysen & Vogel, 2009). In one academic year, 38% of professors surveyed perceived an act of bias in their classes (Boysen & Vogel, 2009). In small group learning classrooms, both explicit and implicit biases have been shown to manifest in many ways, including in peer assessments (Wayland, Walker, & Ferrara, 2014).

Biased actions may result from the implicit and explicit biased attitudes of an individual. These attitudes create an individual's subjective organizational structure for how they perceive their environment (Dovidio, Kawakami, & Gaertner, 2002). Although biased attitudes are deeply engrained, they tend to be inconsistently expressed, depending on the social context. Explicit attitudes in particular are often moderated or "censored" in sensitive social situations (e.g. in public or in a peer assessment) (Dovidio & Fazio, 1992).

In the classroom, instructors perceive incidents of implicit and explicit bias as occurring at similar frequencies (Boysen & Vogel, 2009). Additionally, women and younger faculty members have been shown to be more likely to detect and report biased incidents (Boysen & Vogel, 2009). Classroom bias tends to target individuals' sexual orientation, race, sex, and ethnicity (Boysen, Vogel, Vope, & Hubbard, 2009).

Implicit Attitudes

An implicit attitude is one that is "…activated by the mere presence (actual or symbolic) of the attitude object and commonly function without a person's full awareness or control" (Dovidio, et al., 2002, p. 62). In this context, an "attitude object" is the target of the biased

attitude. Due to this lack of awareness, implicit attitudes are inaccessible through personal introspection (Staats, Capatosto, Wright, & Contractor, 2014). These attitudes shape not only the actions an individual takes, but can influence non-verbal behaviors, such as body language (Dovidio et al., 2002). In short, it is difficult for an individual to recognize and address their own implicit biases without outside intervention.

In the classroom, implicit bias manifests in varying ways. One study of bias in university classrooms divided occurrences into categories of microaggressions. It was found that microassaults (exclusion), microinsults (subtle verbal snubs largely unknown to the perpetrator), and microinvalidations (negating the experiences of marginalized groups) were the most common manifestations of implicit biases in the surveyed classrooms (Boysen & Vogel, 2009). A study by Edith Samuel (2004) produced similar findings on the microaggressions experienced in the classroom. The following examples illustrate types of language and behavior that may indicate microaggressions and implicit bias:

- Microassault: Ignoring the contribution of a group member
- Microinsult: Asking a student of color "How did you get accepted here?" or "But where did you *really* come from?"
- Microinvalidation: Proclaiming to a student of color "I don't see race" or denying any personal implicit biases

Implicit bias is also evident in unstated assumptions about other students such as needlessly offering assistance or doubting a student's ability to complete a task (Samuel, 2004; Boysen & Vogel, 2009). Additionally, manifestation of implicit bias may be observed in body language – such as an inability to make or maintain eye contact, or physically distancing oneself from a member of a disadvantaged group (Chen & Bargh, 1997).

Unfortunately, implicit bias is not reserved for student-student interactions. This type of bias has also been seen in the marks given to students by instructors. Female students have been shown to receive lower class participation scores than male students, despite no evidence for this disparity in other aspects of the course (e.g. exam and homework scores) (Dancer & Kamvounias, 2005).

Explicit Attitudes

An explicit attitude is one that is consciously held about a person or group. These attitudes shape responses for which individuals have the opportunity to consider the social costs and benefits of a particular action (e.g. using a homophobic slur) (Dovidio et al., 2002; Wilson, Lindsey & Schooler, 2000). Where implicit attitudes are difficult to self-recognize, control, and measure, explicit attitudes are overt and more readily measured by traditional assessment measures.

A study on the experiences of South Asian students in predominantly white classrooms found that, out of the 40 students interviewed, all had experienced an incident of explicit bias. Incidences of explicit bias included: openly mocking a student's manner of dress or religious expression (such as the dastaar or hijab), asking rude or inflammatory questions (such as "Did you live in a hut?"), and publicly ridiculing students' abilities and accents (Samuel, 2004). From an instructional perspective, the incidences of explicit bias most reported by professors are the explicit use of stereotypes, telling offensive jokes, and using slurs (Boysen & Vogel, 2009).

Bias in Peer Assessment

Brindley & Scoffield (1998) found that a common apprehension held by students about their participation in peer assessment was that it would be "difficult to avoid personal bias"(p. 85). Bias in peer assessment can be directed at many different social and demographic groups.

While the biases discussed hitherto have focused on gender, race, and ethnicity, peer assessment bias has also been observed due to social style, socioeconomic status, native language, and peer group affiliation.

Bias due to gender has been studied extensively. However, the results have been mixed for both same sex and opposite sex ratings (Eagly & Carli, 1981; Langan et al., 2005). A study of oral presentation assessments found same sex bias where men rated other men slightly higher, however ratings given by women were unaffected (Langan et al., 2005). Conversely, other researchers found that ratings between members of the same sex were consistently more prone to devaluation than when rating the opposite sex (O'Neill, 1985). Males have been found to award the highest scores to females and females award the highest scores to males (Van-Trieste, 1990; Aryadoust, 2016). Recently, though, a study of the implementation of TBL in general education classes showed that while gender bias in the assessment scores was not observed, women did more work in team activities, suggesting that their extra work was going unrewarded (Wayland, et al., 2014).

Results have again been mixed when analyzing the overall ratings *received* by male and female students. Multiple researchers have found that female students receive lower ratings and fewer positive qualitative comments than their male peers (Pajares & Johnson, 1996; Kaufman, Felder, & Fuller, 2000; Bryan, Krych, Carmichael, Viggiano, & Pawlina, 2005). However, it has also been found that males receive lower ratings than their female peers (Sherrard, Raafat, & Weaver, 1994; May & Gueldenzoph, 2006; Tucker, 2014). The mixed results extend to analyses of the ratings *given* by male and female students. Women have, in some cases, been found to give higher evaluation scores than men (Sherrard, et al., 1994). Conversely, another study found men give higher evaluation scores and women give lower marks (Kaufman, et al., 2000).

Student attitudes toward peer assessment are also inconsistent. A study of Australian undergraduates found no gender differences in satisfaction with the peer assessment process (Gatfield, 1999). However, more recent research has shown that male students report more positive attitudes about peer assessment than female students (Wen & Tsai, 2006; Topping, 2010).

One of the few consistent findings for gender effects relates to ratings given by students to themselves. Female students consistently underestimate themselves (Rees, 2003; Lind, Rekkas, Bui, Lam, Beierle, & Copeland, 2002; Das, 1998), while male students consistently overestimate themselves (Rees, 2003; Lind, et al., 2002).

Many researchers have proposed explanations for these differences in results. Falchikov and Malkin (1997) suggested that they may be due to gender-based communication differences and socialization or the gender association of the task or class (i.e. women in a traditionally masculine topic of study). Another explanation is the influence of ability. In many studies where women have received overall higher peer assessment scores, they also had higher GPAs (May & Gueldenzoph, 2006; Baker, 2008). It has been suggested that examining peer assessment scores in conjunction with GPA might clarify this effect (May & Gueldenzoph, 2006). More recently, the mixed results have been suggested to be affected by cultural differences unaccounted for in analysis (Aryadoust, 2016). As the association of behaviors or personality traits with gender varies by culture, this is a plausible explanation.

Studies of racial bias in peer assessment have also returned mixed results. Multiple researchers have found that individuals tend to be rated higher by members of their own race than of other races (Cox & Krumboltz, 1958; Dejung & Kaplan, 1962; Landy & Farr, 1980; Kraiger & Ford, 1985). However, other work has demonstrated that this may not be the case. In

an extensive re-analysis of Kraiger and Ford's (1985) military data, Black raters were shown to give higher ratings to White ratees than to Black ratees (Sackett & DuBois, 1991). Further analysis of these data indicated that Black recruits consistently received lower ratings than White recruits from both Black and White raters (Sackett & DuBois, 1991). Similarly, peer ratings in a sophomore level engineering class demonstrated that minority students received lower ratings than non-minority students (Kaufman, Felder, & Fuller, 1999). Finally, in some cases, no significant evidence of racial bias has been produced (Schmidt & Johnson, 1973; Pulakos, White, Oppler, & Borman, 1989).

Convincing evidence of racial bias was found in a study of peer assessment in large general education classes taught using TBL (Wayland, et al., 2014). In these classes, students of color contributed the same number of answers and suggestions as their peers. However, they received significantly lower peer evaluation scores than White students on three out of four areas of assessment (Wayland, et al., 2014).

The issue of "friendship bias" in peer assessments is familiar to many instructors. Friendship itself has shown to positively impact team functioning in terms of individual accountability and a general sense of community among teammates (Wang & Imbrie, 2010). Additionally, "friend pairs" have shown better ability to constructively critique and develop peers' ideas when tackling challenging tasks (Tolmie, Topping, Christie, Donaldson, Howe, Jessimen, Livingston, & Thurston, 2010). However, concern that these relationships lead to biased peer assessments has been expressed anecdotally by instructors as well as in higher education literature for over 30 years (e.g. Montgomery, 1986; Archer, 1992; Dancer & Dancer, 1992; Rafiq & Fullerton, 1996; Strijbos, Ochoa, Sluijsmans, Segers, & Tillema, 2009). The difficulty of assessing "friendship" within a team, however, has led to a dearth of empirical

studies into the actual extent of this bias. One study sought to investigate reciprocity bias by positing that if Student A rates Student B higher than expected, Student B will also rate Student A higher than expected. This effect was found to be very small (1% of explained variance) and did not account for the actual interpersonal relationships between students (Magin, 2001). Students tend to express reluctance to assess friends harshly or fear that other students may have already favored their friends in evaluations (Cheng & Warren, 1997; Smith, Cooper, & Lancaster, 2002). Instructors have expressed similar concerns that friendship could reduce the validity and reliability of their peer assessments (Karaca, 2009). Similar to bias due to friendship are biases due to collusion and decibel marking. In cases of collusion, students agree to rate each other in certain ways, generally to collectively give inflated scores (Magin, 2001). Collusion is characterized by similar or identical evaluations among teammates (Matthews, 1994). Decibel marking results in higher ratings being given to more dominant group members (Harris & Brown, 2013).

Other factors affecting the fairness of peer assessment have received less attention than gender, race, and friendship. A survey of 232 undergraduates in the United States revealed that 32% would evaluate oral presentations by less wealthy students more harshly than they would evaluate more wealthy students (Moorman & Wicks-Smith, 2012). This effect was particularly strong for female raters. Additionally, students who held more conservative attitudes were more likely to give harsher ratings to less-wealthy students. Bias due to social style has been observed in both engineering and business classrooms with students who exhibit an “expressive” social style receiving significantly higher marks than students with other social styles (Valencia, Carrillo, & Benitez, 2012; May & Gueldenzoph, 2006). Evidence has also suggested that peer

ratings may be biased by language similarity (native speakers receiving higher ratings than non-native speakers) (Langan, et al., 2005).

Benefit and Contributions

Students tend to view the peer assessment process as theoretically beneficial to their learning, an attitude which is echoed in the literature (Thondhlana and Belluigi, 2016). Further, students also find the process of including peer assessment in a grade to be fair *in principle (*Thondhlana and Belluigi, 2016). However, students worry that in practice their peers are not impartial raters, and indeed, research has shown that they are not (Thondlana and Belluigi, 2016; Planas Llado, et al., 2014; Moorman and Wicks-Smith, 2012.) Many of the oft-cited positive outcomes of team learning and peer assessment, such as increased confidence, increased quality of work, and personal accountability are tainted by perceptions of unfairness (Cestone, Levine, and Lane, 2008). Therefore, unfairness in peer assessment dampens the positive outcomes associated with the process for many learners (Thondhlana and Belluigi, 2016).

In engineering, the process of peer assessment is suggested as a method of teaching and reinforcing core professional skills (e.g. communication, peer work review, and team skills). However, negative experiences with teaming and peer assessment could leave engineering students professionally unprepared (Kim, 2014). Calls for increases in engineering education often focus on the diverse talent needed to engage in solving complex socio-technical problems (e.g. improving access to clean water) (Belanger, Diekman, and Steinberg, 2017; NAE, 2015). Unfortunately, the chilly classroom climate some engineering students encounter, which can include bias in team learning and unfairness in peer assessments, can deter diverse students from continuing their engineering education (Gunter and Stamback, 2005).

By studying bias in peer assessment, researchers are able to more fully understand where bias occurs and how it affects students. In turn, this understanding helps with the design of classroom peer assessment trainings, making these efforts more targeted and effective at improving the fairness of peer assessments. If these interventions are successful, the benefits of team learning and peer assessment, such as increased confidence in one's work, better understanding of the material, and improved professional skills may be more equally shared among students, regardless of minority status. Additionally, the implementation of this training could improve the team and classroom climate for minority students in engineering, which in turn could contribute to more of these students completing their studies.

The issues in peer assessment and the expected benefits from resolving those issues give rise to four related research questions:

RQ1. How do students and instructors perceive bias in peer evaluations?

RQ2. What evidence of bias is present in peer evaluation data?

RQ3. What are the requirements of and barriers related to implementing peer evaluation bias training in the classroom?

RQ4. Does bias mitigation training positively impact student perceptions of peer assessment fairness?

Approach

The approach to addressing this problem is divided into three phases which correspond to different parts of the work. The goal of Phase 1 is to clarify the problem of bias in peer assessment (RQ1 & RQ2). There is a gap in the literature for work of this type that spans multiple perspectives (e.g. student and instructor) as well as across departments and classes. This is accomplished through literature review and three studies: a surveys of students on perceptions

of peer assessment bias, a survey of and instructors on perceptions of peer assessment bias, and an analysis of over twenty thousand peer assessment ratings given and received within the Thinkspace peer assessment platform. In Phase 2, the bias mitigation training begins to take shape (RQ3 & RQ4). Using the lessons learned from an initial in-class pilot of bias mitigation methods as well as the results the studies in Phase 1, the first version of the online training is developed. This training then undergoes a formative assessment in a limited number of classrooms (RQ4) and used as a starting point for gathering requirements from stakeholders (RQ3). In Phase 3, the training is further refined according to the revised requirements developed in Phase 2. The refined training then undergoes a summative assessment and is packaged for dissemination and further use.

Document Structure

Six studies are described in this work. After Chapter 2 (Literature Review) provides background research for all studies, Chapter 3 (Approach) will present an overview and approach taken. Next, Chapter 4 will cover the Methods, Results, and Discussion for the first three studies (a survey of students, a survey of instructors, and an analysis of a large body of peer assessment data) on the occurrence of peer assessment bias (Phase 1). Similarly, Chapter 5 (Phase 2) will describe the initial development and formative evaluation of peer assessment bias training. Chapter 6 (Phase 2) describes three focus groups (written together as one study) conducted to assess the requirements for training and a deeper understanding of biased peer assessments. Chapter 7 (Phase 3) will describe a study covering the development and summative evaluation of the final peer assessment bias training. The document will conclude with Chapter 8 and a discussion of what was learned and a review of the contributions made.

CHAPTER 2: LITERATURE REVIEW

This chapter will address the relevant literature to support the research questions developed in Chapter 1. The literature is organized into according to four themes: Peer Assessment support for ABET and student outcomes (related to RQ3 and RQ4), peer assessment benefits and methods (related to RQ3 and RQ4), bias measurement and mitigation techniques (related to RQ1 and RQ2), and challenges in anti-bias training design (related to RQ3 and RQ4). These themes provide the necessary background knowledge required to begin addressing the issue of bias in peer assessment outlined in Chapter 1. The themes are organized in a "top-down" approach. First, standards for engineering education are presented since peer assessment is prescribed as a means to address ABET-specified student outcomes. Next, peer assessment itself will be described to demonstrate the implementation and benefit of these methods in the classroom. As already established in Chapter 1, classroom teams and peer assessments can fall victim to bias. While peer assessment bias has been less studied, there is more work on types of biases and general types of bias mitigation. Therefore, the next section will describe how types of bias in general are measured and mitigated. After understanding specific general bias mitigation strategies, it is necessary to know how to adapt and design them to address specific bias in peer assessment. Therefore, the final section will present challenges in anti-bias training design that must be addressed when developing the specific peer assessment training work of this dissertation.

Peer Assessment to Support ABET and Student Outcomes

The Accreditation Board for Engineering and Technology (ABET) is a nonprofit organization which accredits university programs in engineering, computing, and applied and

natural science (ABET, 2017). Most R1 (Carnegie Classification "Doctoral/Very High Research Activity") engineering programs hold ABET accreditation, which must be periodically renewed by presenting data on student outcomes (ABET, 2017; Indiana University Center for Postsecondary Research, 2021). To ensure quality and consistency, ABET issues a set of criteria which programs must meet. New criteria requires scrutiny on how diversity and inclusion are supported in team work. One method suggested as a measurement of attainment of this criteria is peer assessment. These criteria often undergo minor revisions, however major changes were made to Criteria 3 in the 2019-2020 accreditation cycle.

These changes included more detailed definitions and measureable outcomes and replaced the previous student outcomes (a)-(k) with 1-7. The new Criterion 3 and the associated list of student outcomes is below.

Criterion 3. Student Outcomes: The program must have documented student outcomes that support the program educational objectives. Attainment of these outcomes prepares graduates to enter the professional practice of engineering. Student outcomes are outcomes (1) through (7), plus any additional outcomes that may be articulated by the program.

1. An ability to identify, formulate, and solve complex engineering problems by applying principles of engineering, science, and mathematics
2. An ability to apply engineering design to produce solutions that meet specified needs with consideration of public health, safety, and welfare, as well as global, cultural, social, environmental, and economic factors
3. An ability to communicate effectively with a range of audiences

4. An ability to recognize ethical and professional responsibilities in engineering situations and make informed judgments, which must consider the impact of engineering solutions in global, economic, environmental, and societal contexts
5. An ability to function effectively on a team whose members together provide leadership, create a collaborative and inclusive environment, establish goals, plan tasks, and meet objectives
6. An ability to develop and conduct appropriate experimentation, analyze and interpret data, and use engineering judgment to draw conclusions
7. An ability to acquire and apply new knowledge as needed, using appropriate learning strategies.

While the previous version of the student outcomes addressed teaming (Outcome *d: an ability to function on multidisciplinary teams),* the current iteration (Outcome 5) gives much more detail about the expectation of the team environment including the following definition of a team: "consisting of more than one person working toward a common goal and including individuals of diverse backgrounds, skills, or perspectives" (ABET, 2021). ABET requires measured evidence that outcomes are attained, and with this more detailed definition of the team environment, provides suggestions for how this evidence can be obtained. One specific suggestion is the use of peer assessment: "Use of web-based peer evaluations such as CATME.org or TEAMMATES. The peer evaluations include specific questions about collaboration and inclusiveness." (ABET, 2019, pp. 5). Student Outcome 5 also specifically prescribes that the team environment be inclusive. As outlined in Chapter 1, teaming and peer assessment is not always an inclusive process for all students. Therefore, this work on improving

the fairness of peer assessments can be directly applied to increasing the fidelity of the assessment of ABET Student Outcome 5.

Peer Assessment

Peer Assessment may be defined as a process whereby students evaluate or are evaluated by their peers (e.g. grading a classmate's lab report, providing feedback on a presentation, or scoring a peer's contribution to a team) (van Zundert, Sluijsmand, and van Merrienboer, 2010). Many acedemic and professional products may be assessed in this manner including portfolios, writing, oral presentations, or team performance (Topping, 2009). While peer assessment has become a common topic of study within the last 30 years, it has actually been in documented use since at least the late 18th Century (Gaillet, 1992; Topping, 2009).

The utility of peer assessment has been demonstrated in classrooms from elementary school through higher education and professional school, and been shown to benefit a wide range of students, including those with learning disabilities (Scruggs and Mastropieri, 1998; Topping 2009). Within higher education, peer assessment has become commonplace in nearly all disciplines including: medicine, engineering, humanities, business, education, science, and social science (Falchikov and Goldfinch, 2000). Peer assessment's rise in popularity has been driven by changes in higher education practices, including greater use of teams and collaborative learning, emphasis on practical professional skills, and teaching methods which focus on constructing rather than simply reproducing knowledge (van Hattum-Janssen and Lourenco, 2006; Van den Berg, Admiraal, and Pilot, 2006). Peer assessments are found to be more reliable and better accepted by students when they are supported by peer assessment training, examples, instructor assistance, instructor monitoring, and joint student-teacher construction of the assessment criteria (MacArthur, Schwartz, and Graham, 1991; Topping, 2009).

Benefits of Peer Assessment

When thinking of peer assessment, the first thought that likely comes to mind is "feedback". In the peer assessment process, learners receive plentiful feedback. As there are nearly always more students than teachers, empowering students to give feedback means that students receive more immediate and individualized information (Topping, 2009). Additionally, as students are on the same "level" of authority with each other, the feedback process between them can become a conversation much more easily than between a student and teacher (Cole, 1991). Further, students tend to find this judgment by peers more motivating than instructor assessment (Searby and Ewers, 1997). In a purely practical sense, utilizing peer assessment can also reduce teacher workload over time (Topping, 2009; Falchikov, 2001).

Peer assessment facilitates the development of a variety of professional skills particularly relevant to the higher education classroom, including: interpersonal communication, teamwork, problem solving, organizational skills, and the ability to give and receive feedback (Brindley and Scoffield, 1998; Boud, Cohen, and Sampson, 1999; Woolhouse, 1999; Levine, 2008; Cestone, et al., 2008). This ability to interpret the work of future colleagues is considered a necessary part of professional development (Sluijsmans, et al., 2004; Nicol, et al., 2014). Therefore, learning environments which teach and emphasize this ability can be particularly beneficial for students (Langan, et al., 2005). Many of the skills fostered by the use of peer assessment correspond to those in ABET's Student Outcome 5 (e.g. collaboration, teamwork, goal setting).

From a cognitive perspective, the use of peer assessment has been shown to facilitate deeper learning and improve students understanding of complex topics (Langan, et al., 2005). Peer assessment also allows students to identify errors and misconceptions earlier, which in turn leads to earlier identification of knowledge gaps (Topping, 2009). Further, the act of peer

assessment encourages students to take a more active role in managing their learning (Liu & Carless, 2006). By becoming the assessors, students are required to show a deeper understanding of the material (Searby and Ewers, 1997). Many students, especially in higher education, develop a "superficial" approach to their learning in which they retain just enough information to pass examinations (Thompson and Falchikov, 1998). Peer assessment can contribute to a disruption of this pattern by encouraging deeper thinking (Liu and Carless, 2006).

Peer assessment can also function as a motivational tool for enhancing responsibility and accountability. Responsibility may be enhanced due to the student being an active participant in the peer assessment process rather than simply a passive recipient of a grade (van Hattum-Janssen and Lourenco, 2006). Further, in many classrooms which utilize teams, peer assessment is used as an accountability tool to prevent "social loafing" (failing to participate) (Cestone, Levine, and Lane, 2008). The knowledge that students will receive a grade from peer assessments is an incentive for them to think critically about their contribution to team dynamics and the group's success overall (Michaelsen, 1992).

Types of Peer Assessment and Peer Assessment Methods

Peer assessment often takes one of four forms: formative feedback, peer grading, peer assessment of group work participation, or summative assessment. The use of formative feedback is often used in conjunction with assignment drafts before submitting a final product. Students receive feedback after each draft which can then be applied to future drafts in order to improve the work over time. In this type of peer assessment, students only provide feedback and not a final grade, which is given by the instructor or later via summative assessment. This type of feedback is often conducted through a learning management system such as Canvas or Blackboard, but can also be conducted as an in-class activity (Kennell, Elliot, & Weirick, 2017).

The process of summative assessment is similar to that of formative feedback, but with the intention that the feedback will constitute a grade. Summative assessment nearly always involves the use of a grading rubric to guide the students. Generally, each student is assessed by more than one peer and the final grade is determined by the mean or median of the peer score (Kennell, Elliot, & Weirick, 2017).

In the peer grading type of peer assessment, students assign grades to their peers based upon an assessment rubric. This is often done using online tools (e.g. Canvas) that randomly and anonymously distribute assignments to be reviewed. Students usually give and receive peer grades from multiple classmates, which are then tallied for the instructor to determine the final grade on the assignment (Kennell, Elliot, & Weirick, 2017)..

Instructors may find it challenging to grade group work or team dynamics. The use of peer assessment of group work participation addresses this challenge by placing the assessment in the hands of those who know the group's functioning most intimately – the group itself (Topping, 2009). This type of peer assessment is often implemented as a supplement to grades given by the instructor by adding a participation or team maintenance component to scores of group work. In this method, students score the participation of each teammate and provide feedback on performance. The instructor then uses these assessments to assign overall participation grades (Topping, 2009).

While each type of peer assessment may be implemented in many ways, there are multiple recognized methods of assessing group work in particular. Though not an exhaustive list, the attributes of two such methods will be outlined here.

Michaelsen (balance) and Fink Methods

In the Michaelsen method (also referred to as the balance method), students assign a numerical score to their teammates based upon the extent to which they believe their teammates contributed to the team as a whole (Michaelsen, 2002). Typically, the number of points which can be divided among the teammates is equal to [Number on Team – 1 (students do not assess themselves)] x 10 points (Levine, 2008). For example, if a team had five members, a student completing the peer assessment would be given 40 points to distribute among their four teammates. A hallmark of this method, however, is that students may not assign everyone the same score. For example, if a student was dividing 40 points among four teammates, they would not be allowed to give each teammate 10 points. This forced differentiation encourages students to think carefully and critically about the contributions of each team member (Cestone, et al., 2008). Unlike the Michaelsen method, students using the Fink method are not required to differentiate the number of points given to each team member (Cestone, et al., 2008).

The overall score for each team member is then calculated by summing the scores from each of their teammates. In this method, students also have the ability (which many instructors require) to give both positive and critical qualitative comments (Cestone, et al., 2008). The Michaelsen method can be implemented without specific tools, but Thinkspace and OpenTBL are two online tools which focus specifically on implementing this method.

Comprehensive Assessment for Team-Member Effectiveness (CATME)

Comprehensive Assessment for Team-Member Effectiveness (CATME) is both a method of peer assessment and an online peer assessment tool. In the CATME system, students rate their teammates and themselves on five important dimensions of team-member contributions: contributing to the team's work, interacting with teammates, keeping the team on track,

expecting quality, and having relevant knowledge, skills, and abilities (Ohland, Loughry, Woehr, Bullard, Felder, Finelli, Layton, Pomeranz, and Schmucker, 2012). When evaluating on each dimension, students are shown a set of behavioral descriptions that correspond to low, medium, and high performance on that dimension (Loughry, Ohland, and Woehr, 2013). Students are also able to leave qualitative comments for the instructor and their teammates. A unique attribute of the CATME system is its internal analysis of the peer assessments. This analysis flags the instructor when unusual ratings are detected (high performers, low performers, under/over confident students, and those trying to manipulate the rating system). CATME can also flag team-level concerns such as ratings which suggest internal conflict (Loughry, et al., 2013). This method also features built-in rater training in which students are presented with a fictitious teammate which they must assess. Students then receive feedback about how their ratings compare to the correct ratings for that teammate as well as justification for the ratings (Loughry, et al., 2013; Ohland, et al., 2012).

Bias Measurement and Mitigation

In order to detect and address bias, it needs to be identified and measured. Furthermore, bias mitigation strategies have been implemented in a wide variety of domains and applied in many contexts. This section will describe the techniques used to measure and mitigate bias. Like bias itself, measures of bias are often divided into implicit and explicit measures. Implicit bias tends to be measured through automatic cognitive processes while explicit bias is often measured through self-report questionnaires.

Implicit Bias Measurement Techniques

Measures of implicit bias are generally focused on those which reflect the automatic responses of the cognitive system (Maass, et al., 2000). The most common method of measuring

implicit bias is through the use of the Implicit Association Test (IAT). Developed in 1998 at the University of Washington, the IAT measures association of target concepts with an attribute (Greenwald, McGhee, & Schwartz, 1998). Two concepts (such as Black vs. White) appear in a first task followed by two attributes (such as pleasant vs. unpleasant) in a second task. When the combination of concept + attribute is highly associated, participant's categorization of stimuli items will be quicker than when the combination of concept + attribute is less associated. The difference in performance measures among the different combinations of items measures the implicit association of the concept with the attribute (Greenwald, et al., 1998). The IAT has also facilitated an understanding of the pervasiveness of implicit bias. Over thousands of tests, researchers have found that 80% of Whites and 40% of Blacks harbor pro-White bias (Wald, 2014). Depending on the latency in response time and frequency of errors, the IAT measures the strength of association of each pairing such that more strongly associated categories are easier to pair, reflected by faster responses and fewer errors. Participants who categorize White faces with positive words more quickly and with fewer errors than when categorizing Black faces have an implicit pro-White bias (Greenwald, et al., 2003). Negative scores of the same degree indicate pro-Black bias. Similar scoring algorithms are used to score the additional IATs developed by Project Implicit: Transgender-Cisgender, Gender-Career, Asian-European, Presidential Popularity, Gay-Straight, Religions, Disabled-Abled, Weapons-Harmless Objects, Young-Old, Fat-Thin, Light Skin-Dark Skin, Gender-Science, Native-White Americans. Each IAT is available to the public via implicit.harvard.edu. Study-specific IAT data collections are available through Project Implicit's consulting services.

Despite widespread acceptance of the IAT's validity, it has come under recent scrutiny. Some researchers posit that there is little meaning in the trends shown by the IAT (e.g. 70% of

test-takers prefer faces with European features rather than African features) because of very weak overall connections between implicit bias and discriminatory behavior (Forscher, Lai, Axt, Ebersole, Herman, Devine, & Nosek, 2019; Connor & Evers, 2020). However, one of the original authors of the IAT and director of Project Implicit, Anthony Greenwald, has responded that the IAT was not developed for the purpose of diagnosing prejudiced *behavior* (Bartlett, 2017).

Explicit Bias Measurement Techniques

Explicit biases are generally measured through direct self-report questionnaires. However, as biased attitudes have become less socially acceptable, it has been suggested that this self-reporting cannot be assumed to be accurate, due to respondents moderating their responses to be more socially desirable (i.e. answering what they "should" say rather than what they believe) (Maass, Castelli, & Arcuri, 2000). The "lie-detector-expectation" procedure has been suggested as a method for circumventing participants' response moderation. Using this technique, participants are informed that their physiological responses to the stimuli (e.g. questions on biased attitudes) will be collected in a subsequent session. However, no subsequent session is necessary; the expectation that a "lie detector" will be administered has been shown to be enough to elicit responses which may be considered more truthful (Riess, Kalle, & Tedeschi, 1981). The use of such a technique has been deemed inappropriate for many studies (Maass et al., 2000). No single measurement instrument for explicit bias has emerged (in contrast with the IAT for measuring implicit bias). Due to the lack of one predominant measure and concerns about the accuracy and applicability of traditional measures (Nevid & McClelland, 2010), some researchers have created study-specific measures for assessing explicit bias (e.g. Pantos & Perkins, 2013 and Devine, et al., 2012). One such study-specific measure is the "Shoulds and

Woulds" scale which measures the extent to which individuals predict they *would* act with more prejudice than they *should* (Devine, et al., 2012). In a different study, explicit bias was measured by having participants judge "speaker trait measures" (believability, credibility, trustworthiness, knowledge, expertise, intelligence, competence, likeability, friendliness, warmth, judgment, persuasiveness, presentation style, and clarity) after listening to simulated courtroom testimonies by physicians with differing accents (Pantos and Perkins, 2013). Despite the rising popularity of study-specific measures, there are a number of validated explicit bias measures still in limited use.

One of the earliest measures of explicit bias is the Marlowe-Crowne Social Desirability Scale (MCSD). The full version of the MCSD consists of a 33-item questionnaire that assesses beliefs and the extent to which the individual moderates these beliefs to be more socially desirable (Crowne and Marlowe, 1960). Other measures focus on attitudes toward specific groups. The widely used Modern Racism Scale (MRS) is a self-report measure of the extent of an individual's racist beliefs (McConahay, 1986). This scale measures abstract ideas (i.e. affirmative action) rather than attitudes toward specific individuals or groups (Snowden, 2005). Similarly, the Modern Homonegativity Scale (MHS) measures contemporary attitudes toward lesbians and gay men (Morrison and Morrison, 2003). The Attitudes Toward Women (ATW) scale measures attitudes towards the roles of women (Spence, Helmreich, & Stapp, 1973). While the MRS, MHS, ATW, and similar "modern" scales have been designed to elicit more true responses by measuring abstract attitudes, the items on the scale become frequently obsolete (Maass, et al., 2000; Twenge, 1997). For example, one item on the 1986 version of the Modern Racism Scale, "Blacks are getting too demanding in their push for equal rights" would be completely obsolete today: participants could easily moderate their response to this question to

increase social desirability (Maass, et al., 2000). This frequent need for modification reduces the usefulness of such scales as the validity of the scale would also need to be re-tested after each iteration (Maass, et al., 2000, Twenge, 1997).

Bias Mitigation Strategies

The general bias mitigation techniques presented here will focus on those that show the most promise to be adapted to peer assessment in a team-based classroom environment. Additionally, as explicitly biased attitudes are often self-moderated choices, focus will be placed on techniques specifically intended to reduce implicit bias.

As implicit bias is inaccessible through personal introspection, it can be difficult to mitigate. One of the leading theories in prejudice intervention design is the "prejudice habit model", which treats bias like a habit which can be broken (Devine, 1989; Devine et al., 2012). This model and its accompanying intervention are thus far the only intervention shown to produce long-term changes in bias (two2 years post-intervention) (Devine, Forscher, Cox, Kaatz, Sheridan, & Carnes, 2017; Devine et al., 2012; Forscher, Mitamura, Dix, Cox, & Devine, 2017).

Devine and colleagues suggest prejudice interventions yield the best outcomes when implemented in three stages: 1) becoming aware of the bias, 2) understanding the consequences of bias, and 3) learning and practicing strategies to reduce bias (Devine, et al., 2012; Wald, 2014). Devine et al (2012) posit that to begin breaking the "habit of bias", one must first be made aware of the bias and feel concern about the real-world consequences of biased actions (Devine and Monteith, 1993; Plant, Devine, Cox, Columb, Miller, Goplen, and Peruche, 2009). Awareness of implicit bias is generally accomplished by presenting evidence of personal implicit bias via the Implicit Association Test (IAT) (Monteigh, Voils, and Ashburn-Nardo, 2001; Lee, Quinn, and Heymann, 2017). Understanding the consequences of bias can take multiple forms

including readings, testimonials, and presentation of the overarching outcomes of biased actions (e.g. reduced economic vitality) (Devine, et al., 2017). In the "learning" phase of the prejudice habit breaking intervention, participants learn how to implement evidence-based strategies to reduce biased actions. Prominent strategies for bias mitigation include: stereotype replacement, counter-stereotype imaging, individuation, perspective taking, increasing positive contact, and commitment to change. Many of these strategies are mutually reinforcing. For example, positive contact with those who exhibit counter-stereotypic traits provide fodder for counter-stereotype imaging as well as an opportunity for individuation. This intervention model was originally designed to combat racial prejudice (Devine et al., 2012) but has been shown to be effective at moderating gender bias in STEMM hiring as well (Devine, et al., 2017; Forscher, et al., 2017).

Stereotype Replacement

This technique involves recognizing a stereotypical response (e.g. "women aren't good at math") and consciously replacing it with a rational, non-biased response (e.g. "male and female math scores show no difference when accounting for math classes taken") (Fine, Wendt, and Carnes, 2014). Stereotype replacement requires significant self-reflection where an individual tries to determine why a stereotypical response occurred in order to determine how a biased response could be avoided in the future (Devine, et al., 2012; Monteith, 1993). Unlike other techniques discussed which can be implemented at any time, stereotype replacement requires that an individual first recognize a response as stereotyped before the technique may be implemented.

Counter-Stereotype Imaging

Counter-Stereotype imaging involves providing an individual with a detailed description or picture of a counter-stereotype or asking them to create such a description themselves. The

counter-stereotype may be abstract (e.g. “female leaders”) or specific (e.g. “Kamala Harris”) (Devine et al., 2012; Blair, Ma, & Lenton, 2001). This method was specifically developed to address concerns with the notion that stereotypes should be suppressed, which has been shown to be counter-productive. Instead of attempting to weaken or suppress stereotypical associations, research on counter-stereotype imaging suggests that suggests that individuals may be able to better achieve the same end-goal by strengthening counter-stereotypic associations (Blair, Ma, & Lenton, 2001) While many of the techniques discussed in this section have been developed specifically in relation to implicit bias, counter-stereotype imaging has been shown to be effective at mitigating both implicit and explicit bias (Blair, et al., 2001).

Individuation

Individuation prevents biased actions by learning specifics about others in order to view them as an individual as opposed to an incidence of a stereotype. Using this technique, the lines between the “in-group” and “out-group” are obscured, leading to fairer behavior (Wilder, 1978). Individuation can be as simple as consciously remembering the names and facial features of an “out-group” individual or as complex as the “getting to know you” activities often used in group learning classroom. Individuation is particularly powerful in that it can reduce implicitly stereotyped responses in even young children (Xiao, Fu, Quinn, Qin, Tanaka, Pascalis, & Lee, 2015).

Perspective Taking

Perspective taking involves imagining oneself as the target of a particular bias and contemplating the resulting psychological experience (Todd, Bodenhausen, Richeson, and Galinsky, 2001). This strategy works by increasing the closeness between the individual and the targeted group (Galinsky and Moskowitz, 2000). Perspective taking also significantly increases

an individual's awareness of inequality and creates more positive face-to-face interactions for out-group members (Todd, Bodenhausen, Richeson, & Galinsky, 2011). This technique has been shown to be very effective in reducing implicit racial bias as measured by pre- and post-intervention IAT scores (Todd, et al., 2001).

Increase Positive Contact

Contact is one of the greatest influences on the strength of biases: higher levels of self-reported contact with a group result in lower implicit and explicit bias against that group (Herek and Capitano, 1996; Burke, Dovidio, Przedworski, Hardeman, Perry, Phelan, and Nelson, 2015). Positive contact is generated through interactions with the stigmatized group over a period of time. While contact over time tends to be the most effective at mitigating bias, even short periods of positive contact (such as guest speakers from targeted groups) have been shown to be effective (Kelley, Chou, Dibble, and Robertson, 2008).

Counter-Productive Strategies

Unlike the strategies discussed above, there are also strategies that people take which are, in actuality, counter-productive. When discussing bias, individuals often default to the notion of being "bias blind" (e.g. color blind, gender blind) or trusting in one's own ability to make objective judgments. These informal strategies are particularly counterproductive (Carnes et al., 2015). Bias blindness, also referred to as stereotype suppression, has been shown to backfire and produce a rebound effect. These suppression efforts often lead to increased stereotypical response that is greater than if no mitigation effort had been taken at all (Monteith, Sherman, & Devine, 1998; Apfelbaum, Sommers, and Norton, 2008; Carnes, et al., 2015). A strong sense of one's own objectivity (the "I think I am objective, therefore it is true" mindset) has also been shown to increase active discrimination (Uhlmann and Cohen, 2007, pg. 207). Specifically, when

individuals are confident that they are objective, rational actors, they are more likely to act on group-based biases (Uhlmann and Cohen, 2007; Carnes et al., 2015).

Challenges in Designing Anti-Bias Training

Research on the impact of bias mitigation training on peer assessments in higher education is limited. However, a study that combined standard peer assessment rater training methods with bias reduction training produced a significant reduction in bias against individuals with differing social styles (May, 2008). In this study, the bias reduction training was limited to an introduction of the concept of social style and suggestions to avoid biased based upon social style. In future work, including the work described in this paper, a more targeted approach to bias mitigation could be taken. While many bias mitigation strategies (e.g. stereotype replacement, individuation, counter-stereotype imaging) strategies can be employed in the classroom, higher education instructors often try to mitigate biased actions in other ways. A study of university professors found that few employed specific strategies and instead relied on using the bias as a discussion topic, providing a rebuttal or counterevidence, and direct confrontation (Boysen and Vogel, 2009). Therefore, more work on anti-bias training for use with peer assessment is needed.

Anti-bias and diversity training can be implemented in many ways, including video games (e. g. Olson and Harrell, 2020), interactive presentations (e.g. Devine, et al., 2012), and even devised theatre (e.g. Iverson and Seher, 2014). However, before such training can be implemented, its content and scope must be created. This section will focus on the challenges associated with anti-bias training that must be addressed before such training can be carried out. While the challenges themselves are general to anti-bias training, examples will be provided of how they may be addressed in the context of peer assessment.

When designing training, the designer(s) must first start with themselves. In practice, this means developing "critical cultural consciousness", or a self-awareness of the interaction of culture, biases, and discriminatory practices as well as understanding of how their personal beliefs can affect their attitudes toward those who will complete the training (Lin, Lake, and Rice, 2008). The training designer should be able to question and examine their own beliefs about others, as well as whether they are able to see all participants as learners regardless of gender, race, class, or other demographic attributes (Cozart, Cudahy, Ndunda, and Van Sickle, 2003; Lin, et al., 2008). In short, the designer should not view themselves as being without bias.

Carter, Onyeador, and Lewis (2020) argue that "diversity training should go beyond telling people that bias exists or creating uncomfortable experiences that are more likely to prompt defensiveness than learning." (pg. 58). They also suggest that to be effective, anti-bias training should be designed to increase awareness of the occurrence and consequences of bias, and teach skills that enable learners to change their behavior accordingly (Carter, Onyeador, and Lewis, 2020), an attitude shared with Devine and colleagues (2012). To design effective, modern, anti-bias training, five challenges should be addressed: setting realistic expectations for what training can accomplish, selecting proper goals, deciding how to manage discomfort, minimizing counter production, and demonstrating impact (Carter, Onyeador, and Lewis, 2020).

Setting Realistic Expectations

In many cases, the administration of bias training is seen as a "one time investment". However, as discussed in Chapter 1, bias is a multifaceted problem which can affect each individual differently. Everyone has been exposed to different environments and internalized bias in differing ways, therefore, training without reinforcement is unlikely to have the desired effect (Payne, Vuletick, and Lundburg, 2017; Carter, et al., 2020).

In order for trainees to benefit the most from the training, it should be accompanied by other practices to combat prejudice. In other words, relying on trainees' goodwill is unlikely to create change (Chang, Milkman, Gromet, Rebele, Massey, Duckworth, and Grant, 2019; Devine, et al., 2012). Other practices to implement alongside training include: create or improve the process for responding to bias, hold individuals accountable for reducing bias (environmental change), actively prioritize diversity and respond to the use of stereotypes, and remind trainees about techniques for reducing biased responses (Carr, Dweck, Pauker, 2012; Devine, et al., 2012). In the peer assessment classroom, these practices would be similar: instructors could inform students of how to report bias, prioritize diversity by presenting work from diverse sources or inviting diverse speakers, and remind students about the techniques from the training at each peer assessment.

Selecting Goals

Training is similar to teaching, and when designing a course, one of the first items which must be completed is to identify learning outcomes. The same is true when designing bias mitigation training (Lai, 2019). A common pitfall of such trainings is to select only one learning outcome: developing an awareness of bias among trainees (Carter, et al., 2020). The choice of simply making trainees *aware* of bias is rooted in research that has shown training can be very effective at teaching about and raising awareness of bias (Bezrukova, Spell, Perry, and Jehn, 2016). However, being aware of bias alone does not necessarily impact an individual's behavior. To change behavior, training needs to include a behavioral component, which is often overlooked (Carter, et al., 2020). In short, bias mitigation training for peer assessment should teach students actionable skills to improve their ability to provide fair assessment and feedback.

This behavioral component is integral to the effectiveness of the training: training that incorporates both an awareness and behavioral component is significantly more effective at changing attitudes and behavior than awareness-only training (Bezrukova, et al., 2016). There are many recognized strategies which can be included in this behavioral component (discussed in the previous section, Bias Mitigation Strategies) including: counter-stereotype imaging, stereotype replacement, individuation, positive contact, and perspective taking.

Managing Discomfort

It is inevitable that conversations about bias will be uncomfortable to some degree. These conversations often prompt discomfort for both those who may be the target of bias and those who may be the perpetrators (Carter, et al., 2020). Discomfort can cause trainees to become defensive or belittle the training content in order to avoid continuing to experience negative emotions (Brannon, Carter, Murdock, Perriera, and Higgenbotham, 2018; DiAngelo, 2011). For those who have been the target of bias, intergroup discussions of bias may be uncomfortable due to anxiety about being seen as "complaining" or "accusing" others. Additionally, being confronted with examples of bias against a minority group to whom the participant belongs can be painful (Schults, Gaither, Urry, and Maddox, 2015; Trawalter and Richeson, 2008). For trainees who may have been the perpetrators of bias (the majority group), recognizing personal bias can be distressing and provoke strong feelings of defensiveness. Some individuals may be reluctant to acknowledge that they (like everyone) harbor biases which may affect their behavior. In extreme cases, potential trainees may be openly hostile toward the training (Brannon, et al., 2018; DiAngelo, 2011). For training to be effective, this discomfort must be managed.

To manage discomfort, the training designer must know their audience in order to understand the content to include and the types of discomfort that may arise (Carter, et al., 2020).

For example, if the training will be delivered to an audience that is primarily people of color, less attention to awareness of bias may be needed, as these individuals often learn about and experience bias from a young age (Brown, Tanner-Smith, Lesane-Brown, and Ezell, 2007). Instead, training which focuses on coping with biased actions, and procedures for reporting bias while maintaining personal and professional safety would likely cause less discomfort and offer more utility. In contrast, training for an audience that is primarily white would likely put emphasis on awareness of bias and its consequences. In this case, discomfort may be diffused by taking the focus off the individual as a perpetrator of bias and framing it as an issue for everyone to address (Jordan, Spencer, and Zanna, 2003). Similar methods for diffusing discomfort would be appropriate for use in peer assessment bias training. However, research has shown that "a moderate amount of discomfort is a critical catalyst for the introspection that can guide a person toward more egalitarian behavior in the future" (Carter, et al., 2020, pp. 63). Facilitators should not endeavor to remove all discomfort, for it is in discomfort that change may flourish.

Minimizing Counterproductive Effects

Conversations about addressing bias can, ironically, be counterproductive. In organizations which explicitly value diversity, claims of bias are more readily dismissed than in organizations that have no stated position on diversity (Kaiser, Major, Jurcevic, Dover, Brady, and Shapiro, 2013). This could be due to the "I think I am objective, therefore it is true" mindset that has been shown to increase active discrimination (Uhlmann and Cohen, 2007, pp. 207). Framing bias as a simple issue that may be addressed through a single training can also decrease empathy for victims of bias and reduce perceptions of severity of biased actions (Ikizer and Blanton, 2016). In contrast, framing bias as all-encompassing can deter participants from undertaking bias-mitigation efforts (Duguid and Thomas-Hunt, 2015).

Counterproductive effects may be minimized by viewing the training as an exercise in persuasion. Classic social psychology research indicates that persuasion is most successful when individuals are presented with a moderately disturbing outcome and strategies to avoid that outcome (Witte, 1992). This tactic can readily be applied to bias training: trainees may be presented with the consequences of bias and then supplied with strategies which will help them change their behavior (Carter, et al., 2020). When designing training specifically to target bias in peer assessment, such consequences could be tailored to the peer assessment environment (e.g. a student may choose to switch majors due to consistently poor peer assessment and team experiences). For this approach to be effective, the strategies presented must be concrete (Blascovick and Mendes, 2000) Research also suggests that trainees not be overwhelmed with strategy options: two to three strategies are ideal for promoting change (Gollwitzer, 1999).

Demonstrating Impact

To determine whether training is successful and to understand ways in which it should be modified, it is necessary for the training to be evaluated. As a teacher assesses whether learning outcomes are met, so should a training facilitator assess whether goals are met (Jefferson and Lewis, 2018). Evaluating the training can take many forms. To determine the acceptance of the training by trainees, a post-training survey may be useful (Carter, et al., 2020). To determine if the training was able to change implicit attitudes, pre- and post- training IATs could be used (e.g. Devine, et al., 2012). Other metrics could also be appropriate, depending on the goals of the training. For example, if the training was designed to increase fairness in peer assessment tracking the perceived fairness of peer assessments over time as well as peer assessment scores would be of interest.

Instructor Adoption

In this chapter and the preceding chapter, the problem of bias in peer assessment was explored, the mechanisms behind bias were outlined, and the design of anti-bias training was discussed. However, for this work to move beyond the page and into the classroom, it must be adopted by instructors. A primary challenge in higher education is the lack of background knowledge in teaching strategies: "few professors have actually been taught how students learn and how to teach their own students" (Knobloch and Ball, 2006, p.4). This may be doubly true for teaching strategies that foster diversity and discourage bias (Considine, Mihalick, Mogi-Hein, Penick-Parks, and Van Auken, 2014). Unfortunately, engineering educators and engineering education researchers tend to have especially limited contact with education and social science theory (NRC, 2012). Instructor adoption of teaching strategies (and trainings) designed to discourage bias and foster diversity may be increased through introducing the importance of the method over time rather than in a workshop or meeting, providing feedback or coaching for instructors adopting a new method into their classroom, and encouraging instructors to engage with the content in order to change their own perceptions of their students, not just students' perceptions of each other (Henderson, Beach, and Finkelstein, 2011; Considine, et al., 2014).

CHAPTER 3: APPROACH

In this chapter, the approach for understanding the prevalence of bias in peer assessment and the development, implementation, and evaluation of a training intervention to address this bias is discussed. In the preceding introduction and literature review, occurrences of bias in peer assessment were documented. This knowledge will be expanded upon by investigating the occurrence of bias through data collection and analysis, and creating an online peer assessment bias mitigation training tool. Gaining a better understanding of where, how, and when bias occurs in peer assessment will lead to a more targeted training. Additionally, the implementation of the training online has the potential to be better accepted by instructors.

Research Questions

The goal of the project is to understand the occurrence of bias in classroom peer evaluations and mitigate this bias through training. This project has four research questions.

RQ1. How do students and instructors perceive bias in peer evaluations?

R2. What evidence of bias is present in peer evaluation data?

R3. What are the requirements of and barriers related to implementing peer evaluation bias training in the classroom?

R4. Does bias mitigation training positively impact student perceptions of peer assessment fairness?

The first research question (RQ1) focuses on attitudes and perceptions of bias in peer assessment, and how these may differ with perspective (student versus instructor). The second research question (RQ2) looks to examine the occurrence of bias empirically through analysis of a large body of peer assessment data. Together, R1 and R2 lead to a better understanding of

where bias happens, to whom bias is directed, and its prevalence. This understanding of the types of bias most prevalent in peer evaluations is important when crafting relevant training materials. Previous work has documented evidence of peer assessment bias in individual classes, however the aim of these research questions is to look for bias across varying disciplines in order to create more generalizable training.

The third research question (RQ3) addresses the functional requirements of bias mitigation training as well as potential barriers to training adoption. Following the user-centered design process, the identification of requirements represents the first step in creating a training which meets the needs of its users. This research question also seeks to understand barriers to participation in training. As willingness to adopt a new method or intervention into the classroom can vary widely, knowledge of potential barriers to adoption and how to overcome them may improve instructors' willingness to use the intervention in their classroom (Monahan, McDaniel, George, & Weist, 2014).

The fourth research question (RQ4) focuses on the efficacy of the training. Students often view the peer assessment process as unfair (e.g. Wayland, et al., 2014) due to the potential for biased ratings or perceived lack of qualification of their peers. Evaluation of the proposed intervention aims to improve perception of peer assessment fairness, which in turn may have positive impacts on student willingness to fully participate in peer evaluation.

Project Approach

As Holmes (2020) notes, researchers should assess how their own positions and experiences might contribute to their interpretations of the experiences of others. As this project has involved interpretation of data relating to biases that the authors have not experienced, it is important to state their positionality. The primary author is a US-born White woman under

faculty supervision of a US-born White man. While unintentional, it is possible the ethnoracial backgrounds of the researchers may have influenced their interpretations of data associated with this project.

Figure 1 illustrates the approach to addressing the five research questions associated with this work. The figure is divided into three Phases which correspond to different parts of the work. The goal of Phase 1 was to explore the problem of bias in peer assessment. This was accomplished through literature review, surveys of students and instructors on perceptions of peer assessment bias, and an analysis of over twenty thousand peer assessment ratings given and received within the Thinkspace peer assessment platform. In Phase 2, the bias mitigation training began to take shape. Using the lessons learned from an initial in-class pilot of bias mitigation methods, an analysis of a body of peer assessment data, as well as the results of surveys on student and instructor perceptions of classroom and peer assessment bias, the first version of the online training was developed. This training was then deployed in a limited number of classrooms and used as a starting point for gathering requirements from stakeholders. In Phase 3, the training was further refined according to the requirements developed in Phase 2. The refined training was then evaluated, and is ready to be packaged for dissemination and further use.

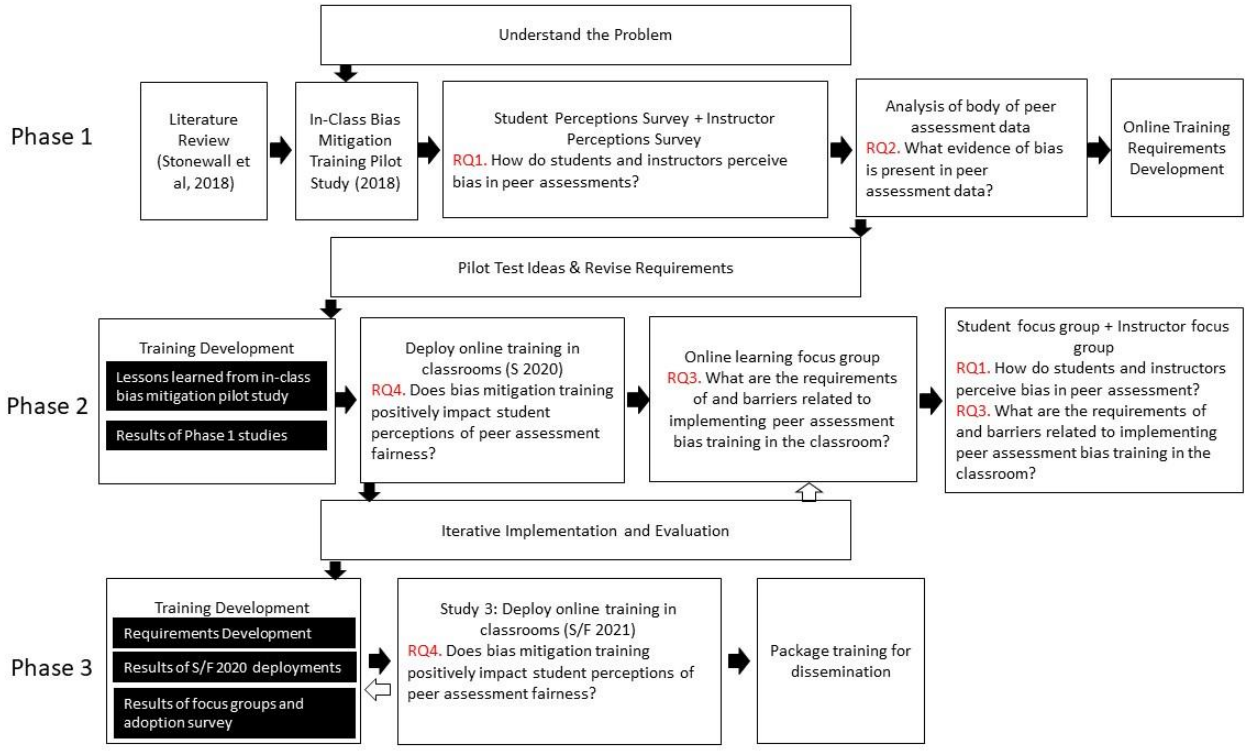

Figure 1. Diagram of Approach

Phase 1: Understanding the Problem

The employment of small group learning strategies (such as cooperative learning, project-based learning, or Team Based Learning) in classroom environments has been shown to increase student achievement, attendance, engagement, and lead to better overall learning outcomes (Michaelson, Knight, & Fing, 2004; Michaelson & Sweet, 2011; Allen, Copeland, Franks, Karimi, McCollum, Riese, & Lin, 2013). Because of these positive outcomes, team-based pedagogies and cooperative learning practices have been incorporated on college campuses as a strategy to improve the classroom engagement of underrepresented students. In many team-based pedagogies and classrooms, peer assessments are integral. However, both students and instructors have expressed concerns about the fairness of teams and their associated peer

assessments, especially due to bias (Magin & Helmore, 2001; Samuel, 2004; Dancer & Kamvounias, 2005; Aryadoust, 2016). In Phase 1, the prevalence of bias in peer assessment was explored. This exploration begun with a broad literature review on the implementation of peer assessment, its strengths and weaknesses, and the ways in which bias may impact peer assessment (Stonewall et al., 2018). Through literature review, biases were identified along the lines of race, gender, and socioeconomic status. These findings motivated further investigation into the prevalence of bias through the use of surveys and data analysis.

Student and Instructor Perceptions Surveys

Previous work on peer assessment bias has focused primarily in individual classrooms or academic departments (e.g. May, 2008; Wayland, et al., 2014). Surveys were conducted with students and instructors to gain understanding of peer assessment bias across classrooms and departments. These surveys were designed and distributed with the intention of filling in a broader picture of the issue of bias by being distributed to a larger audience across a university. Students were asked about their perceptions of fairness and bias within peer assessment, as well as their level of experience with and general attitude toward peer assessment. While perceptions of bias do not necessarily indicate that bias is present in peer assessment scores or grades, it impacts the climate of the classroom and students' acceptance of the peer assessment process (Paquet & Des Marchais, 1998). Instructors received similar questions about the occurrence of bias in their classrooms and their peer assessments. Additionally, instructors were asked about potential actions taken to mitigate bias. Together, these surveys address RQ1 by determining how instructor and students perceive bias in peer assessments. These surveys and their results are discussed in detail in Chapter 4.

Analysis of Peer Assessment Data

The perceptions of bias by students and instructors are only one side of the "bias picture". The other side of the picture is formed by an objective analysis of peer assessment data. In this study, such an analysis was performed on over 20,000 peer assessment ratings given and received by nearly 9,000 students within Thinkspace between 2014 and 2018. These ratings were connected to demographic data, which were then analyzed for evidence of bias. Such evidence was found in terms of gender and race, and address RQ2. Results of this analysis are discussed in detail in Chapter 4.

Phase 2: Pilot Test Ideas and Revise Requirements

To design the training, a user centered design process was utilized. User-centered design (UCD) refers to a process of developing a product for the end-user or the person who will be using the product (Abras, Maloney-Krichmar, & Preece, 2004). Typically, the user-centered design process is an iterative process which involves initially discovering what is needed through requirements gathering, designing a product, implementing, and evaluating the design (Dix et al., 2004). As the process cycles, requirements may be refined or wholly redefined, which then lead to design changes and further evaluation.

Initial Training Development

The initial design of the training was guided by basic requirements gathered from the studies in Phase 1, as well as a previous pilot study involving an in-class peer assessment bias mitigation training. These requirements included: online delivery, content covering biases due to gender, race, age, and socioeconomic status, and content covering how to give helpful feedback, and criteria on which to assess peers. The training was designed in two parts: Part 1 which covers best practices for feedback and evaluation criteria, and Part 2 which covers types of bias, how

bias affects others, and how biases may be mitigated. Material for Part 1 was heavily drawn from the literature on peer assessment, evaluation, and feedback, while the content of Part 2 pulled from literature on "de-biasing" strategies, interventions, and microaggressions (Michaelson, Knight, & Fink, 2004; Michaelson & Sweet, 2011; Cestone, et al., 2008; Devine, et al., 2012) .

Deployment of Training in Classrooms (S 2020)

This initial iteration of the training was then evaluated in a limited number of classrooms during the Spring 2020 semester: four classrooms begun the evaluation, but only one completed the evaluation due to the COVID-19 pandemic. These initial evaluations form part of the response to RQ4. When evaluating the training, students' attitudes toward peer assessment, confidence in others' ability to rate fairly, the fairness of peer assessment overall, and their willingness to participate in peer assessment are measured. Improved attitudes, confidence, and perceptions of overall fairness will indicate a positive response to the training. Peer assessment scores themselves will also be collected and analyzed for both evidence of bias and bias mitigation. The results of this formative evaluation were used to understand the initial performance of the training and identify how it could be improved.

Focus Groups with Online Learning Leaders and Instructors

An evaluation of the requirements for the training was conducted in order to determine how to design training to increase instructor willingness to adopt the intervention into the classroom, as well as to ensure it meets the standards of online learning experts. This evaluation took the form of focus groups with instructors and online learning leadership. These individuals were asked to provide feedback on content, functionality, and method of delivery, which forms the answer to RQ3. This evaluation is further described in Chapter 6.

Focus Groups with Students and Instructors

A deep dive into the perceptions of students and instructors on classroom and peer assessment bias was also conducted in the form of focus groups. This information forms part of the answer to RQ2. Additionally, these focus groups had a secondary focus on feeback on the training. These individuals were be asked to provide feedback on content, functionality, and method of delivery, which form the answer to RQ3. This evaluation is further described in Chapter 6.

Phase 3: Iterative Implementation and Evaluation

Phase 3 is the last stage of implementation and evaluation. The results of the studies in Phase 2 (deployment of online training; online learning, student, and instructor focus groups; instructor adoption survey) were be crafted into requirements which were implemented in the final version of the training. This training was then rigorously evaluated (RQ4).

Summative Evaluation of Leaning Outcomes

The second evaluation of the training proceeded similarly to the first evaluation. Once again, students' attitudes toward peer assessment, confidence in others' ability to rate fairly, the fairness of peer assessment overall, and their willingness to participate in peer assessment were measured. Improved attitudes, confidence, and perceptions of overall fairness indicate a positive response to the training. Peer assessment scores themselves were collected and analyzed for both evidence of bias and bias mitigation. Students who completed the training were interviewed to determine their response to the training as well as retention of training material. The results of iterations of the training and its evaluation inform the response to RQ4.

Packaging the Training

For the training to be useful beyond this project, it is necessary to package it as a standalone "product". The requirements for this packaging were garnered from what was learned

during the student, online learning leader, and instructor focus groups as well as continued work with online learning experts and the results of the evaluations in Phase 3.

Approach Summary

The process of understanding and mitigating peer assessment bias is divided into three Phases. In Phase 1 of the project, the questions of how bias occurs, where bias occurs, and to whom it is directed are answered through literature review, surveys, and analysis of peer assessment data. Next, in Phase 2, an initial bias mitigation training is developed and implemented using requirements from pilot testing and the results of Phase 1. These requirements are then further refined through the use of focus groups before being implemented in Phase 3. Phase 3 represents the final proposed stage of the process. In this phase, the training is refined, implemented, and further evaluated for its effects on student attitudes toward peer assessment, perceptions of fairness, and confidence in others' ability to rate fairly. Finally, using recommendations from online learning experts and instructors, the training will be packaged for further classroom use.

CHAPTER 4. EVALUATION OF BIAS IN PEER ASSESSMENT IN HIGHER EDUCATION

Research Objectives

Some researchers have examined the occurrence of bias in peer assessment in individual classes (e.g. Wayland, et al., 2014) however there is a gap in the literature for a broad investigation of bias across classes, departments, and academic fields as well as analysis of a large body of peer assessment data for bias. Therefore, a series of studies were conducted to understand the occurrence of bias in peer assessment. The studies were designed to look at the issue of bias from three perspectives: instructor, student, and peer assessment data alone.

In the first study, the instructor perspective was explored. Instructors were asked broadly about the occurrence of bias in their classrooms and peer assessments as well as actions taken to mitigate bias. Similarly, in the second study, the student perspective was explored. Students were asked about their experiences with peer assessment and any biases they had perceived. In the third study, a large body of peer assessment data was analyzed to determine what types of bias were present, if any.

Method and Results: Instructor Survey

An Institutional Review Board (IRB) approved survey study (Appendix C, Appendix D) of classroom and peer assessment bias was conducted with Iowa State University instructors. Instructors' responses were recorded and incidence of bias was analyzed.

Participants

Participants for the study were recruited by utilizing the "faculty and staff" email list through Iowa State University Human Resources, distributed through Information Technology Services "big mail" mass email system. The email recruitment message indicated the purpose of

the study as well as its intended audience: instructors. Sixty-one instructors participated in the study.

Out of the instructors who participated in the study, 54 completed the demographics section. Thirty-eight participants identified themselves as women, 11 identified themselves as men, and five preferred not to state a gender. Forty-six participants identified as White, three participants identified as Asian, and five participants identified as "Other". No participants identified as Black, Native American, or Hispanic/Latinx. Forty-eight participants were native English speakers, while six were not. Participants also identified the college with which they were affiliated (Table 1).

Table 1. Participant counts by academic college (N = 54)

College	Participant Count
Agriculture and Life Sciences	9
Business	5
Design	7
Engineering	7
Human Sciences	7
Liberal Arts and Sciences	18
Veterinary Medicine	1

Procedure

Participants began the study by clicking a link to the survey contained in their recruitment email. This link took them to the consent form for the study. After consenting, participants began the survey. A summary of the survey questions and their types may be found in Table 2.

Table 2. Questions in the instructor survey. * indicates a demographic question

Question	Type
Do you use peer assessment in any of the courses you teach?	Yes/no
Are these courses:	Multiple choice: Undergraduate, Graduate, Both

Table 2 Continued	
Question	**Type**
Are any of these courses taught using Team Based Learning?	Yes/No
What types of classes are they?	Multiple choice **(select all that apply):** Seminar, Lecture, Large Lecture, Lab, Studio, Other
On average, how often do you conduct peer assessment?	Multiple choice: Once a semester, twice a semester, 3 times a semester, More than 3 times a semester
What peer assessment method(s) do you use?	Multiple choice **(select all that apply):** Michaelson, CATME, Percentage Effort, Categories with Likert Scale, Qualitative Feedback, Other
Is peer assessment incorporated into the course grade?	Yes/no
Follow-up: How is it weighted or incorporated?	Free response
Does your peer assessment protocol require that students include qualitative comments?	Yes/no
Follow-up: Please describe	Free response
Do you provide students with any training prior to administering peer assessments?	Yes/no
Follow-up: Please describe the training	Free response
Have you perceived any biases in peer evaluations?	Yes/no
Follow-up: Please describe any biases you may have perceived	Free response
Have you taken any steps to mitigate bias in peer evaluations?	Yes/no
Follow-up: Please describe any mitigation strategies you have used	Free response
*What gender do you identify with?	Multiple choice: Woman, man, other, prefer not to say
*How would you describe yourself?	Multiple choice **(select all that apply):** White, Black or African American, Native American or Alaska Native, Asian, Native Hawaiian or Pacific Islander, Hispanic or Latinx, Other
*Are you a native English speaker?	Yes/no
*Which college are you affiliated with?	Agriculture and Life Sciences, Business, Design, Engineering, Graduate, Human Sciences, Liberal Arts and Sciences, Veterinary Medicine
*Which department(s) are you affiliated with?	Free response

Results: Instructor Survey

Survey data was analyzed for counts and themes. Results will be presented in the same order as the questions in Table 2.

Course Information

Each participant provided information about the courses they teach and their implementation of peer assessment. These results are presented in Table 3.

Table 3. Course information provided by instructors

Item	Result
Course Level	Undergraduate: 38; Graduate: 7; Both: 11
Taught using Team Based Learning?	Yes: 20; No: 36
Course Type	Lecture: 34; Seminar:12; Large Lecture: 11; Lab: 10; Other: 9; Studio: 6
Peer Assessment Frequency (per semester)	Once: 17; Twice: 14; Three times: 12; More than three: 13
Peer Assessment Method	Qualitative feedback: 34; Categories w/Likert scale: 23; Other: 9; Percentage effort: 8; Michaelsen: 6; CATME: 6
Peer assessment incorporated into the course grade?	Yes: 41; No: 14 The most common method of incorporation was through setting peer assessment to a percentage of the final grade.
Qualitative feedback required?	Yes: 43; No: 12 Many instructors who did not require qualitative feedback indicated that it was strongly encouraged or only required if giving a negative evaluation. Of those instructors requiring qualitative feedback, the most common implementation was to require at least one piece of feedback on strengths and at least one piece of feedback on areas which could be improved.
Provide students with any training prior to administering peer assessments?	Yes: 32; No: 23 For those instructors who used training, the most common methods were: the built-in training in CATME, providing examples of helpful and non-helpful feedback, and completing practice or sample feedback prior to the first peer evaluation.

Perceptions of Bias

In the survey results, 47.3% of participants perceived bias in their peer evaluations and 52.7% did not. Of the 29 participants that indicated perceived bias, they noted bias due to gender (10 mentions), race (6 mentions), interpersonal relationships (6 mentions), language (3 mentions), and "gaming the system" (5 mentions). Selected quotes in Table 4 illustrate the range of issues instructors mentioned in their replies.

Table 4. Range of issues instructors mentioned in their replies

Gender
"Females are more likely to deprecate their work in teams."
"I have perceived sexist biases against women"
"I've actually been very impressed by the maturity of the student responses but there are issues of women being evaluated differently than men that I've noticed."
"I've noticed that the ways in which feedback is given varies based on gender - women tend to be less direct unless they are giving feedback anonymously"
Race, Ethnicity, Country of Origin
"Students of color have raised concerns that they are not being treated fairly"
"I have had students from China tell me they are very uncomfortable giving a negative evaluation."
"I have seen white students think that Chinese students were not taking the class seriously by not participating when in actuality the Chinese students were struggling with the language."
Interpersonal Relationships
"…students not feeling comfortable saying negative things about their peers, especially bullies- so they receive good feedback."
"Some members seem to give higher ratings to peers automatically so as not to hurt feelings or to "be nice." Conversely, some students who have personality conflicts with a peer seem to give ratings that are much lower than what other peers give the same student."
"We are a cohort based program so I've had issues of students grading assessments of their peers whom they dislike"
Non-Compliance/ "Gaming"
"Students don't report actual peer performance. They're biased against the feedback form."
"I have found in both graduate and undergraduate courses that students tend to be very uncritical of their peers. They argue that they know how hard it is to do the work, so they should be praised for doing it rather than have a list of comments and a rating that shows where improvement is needed."
"Students tend not to want to honestly peer assess their team members even if they complain that the student is not participating effectively in the team. I do not carry out peer-assessment for that reason now."
Other Biases
"Lower SES or non-traditional students receive lower grades."

Comments on gender bias focused on biases women face from their peers as well as through self-deprecation. Instructors also noted that many students are uncomfortable giving

constructive feedback based upon cultural norms, interpersonal relationships, or a lack of compliance with the feedback process.

When asked if they had taken steps to mitigate bias, 51.9% said they had, while 48.1% had not. Of those participants who had taken steps to mitigate bias, common strategies included: class discussion of appropriate evaluation criteria (10 mentions), the instructor "checking" the evaluations (7 mentions), and class discussion of sources of bias (5 mentions). Selected quotes from instructor responses are presented in Table 5.

Table 5. Selected quotes from instructor responses on methods of mitigating bias

Methods of Mitigating Bias – Selected Quotes
"For the qualitative responses we do discuss that assessment should be balanced and informative. Specifically I discuss assessment as a source of information for the student being assessed rather than a positive or negative evaluation."
"Last spring, at the outset of class, I asked each team to anticipate the types of challenges they might encounter in the context of teams where I have aimed to maximize for diversity (e.g., gender, farm background or not, ethnicity, race, technology, sustainability, etc.) and asked each team to compose a team compact by consensus and then we share those statements and discuss as a whole class; before and after rounds of Peer Evals, I talk about what constitutes constructive comments"
"I've stated that peer assessments shouldn't be based on liking, but rather on actual work contributed."
"I discuss sources of potential bias and ask students to focus on the evaluation without bias."
"I try to set up my teams in order to minimize bias (make sure teams are gender balanced and that there is either zero or more than one person of color on teams to avoid someone being a "token""
"I use the output from CATME to identify outlier ratings, like those that would occur from personal conflicts or attempts to manipulate the ratings. When these kinds of behaviors are flagged in CATME, I follow-up by holding individual conferences with each student involved to better understand the ratings that were assigned."

Instructors mitigated bias through team selection to balance and maximize diversity. Instructors also framed the evaluations as informative while emphasizing the purpose and content of good feedback.

Method and Results: Student Survey

An Institutional Review Board (IRB) approved survey study (19-516) of classroom and peer assessment bias was conducted with Iowa State University students. Participants for the study were recruited by utilizing the student email list through the Iowa State University Office

of the Registrar, and distributed through Information Technology Services “big mail” mass email system. The email recruitment message indicated the purpose of the study as well as its intended audience: students who had experienced peer assessment and were age 18 and older.

Participants

This part of the study included 419 participants. Two participants had not worked with a classroom team and were excluded, and 77 participants completed less than half the survey and were excluded. As a result, the study included data from the remaining 342 participants (203 women, 133 men, 5 no answer, 1 other). Participants ranged in age from 18-55, with a median age of 20. Two-hundred ninety-five participants were white, 20 were Asian, 16 were Hispanic or Latinx, six were Black or African American, three were another race or ethnicity, two were Indigenous American or Pacific Islander.

For 321 participants, English was the first language they learned to speak, while 19 participants initially spoke another language. For those participants who did not initially speak English, the mean speaking time was 18.1 years ($n = 18$, $SD = 4.1$ years). Participant college affiliations were: Engineering (110), Liberal Arts and Sciences (75), Agriculture and Life Sciences (58), Human Sciences (30), Business (30), Design (27), Graduate (6), and Veterinary Medicine (6). All class levels were represented in the study: 66 freshmen, 63 sophomores, 67 juniors, 89 seniors, and 54 graduate students. Seventeen participants reported being international students while 325 were not.

Procedure

Participants began the study by clicking a link to the survey contained in their recruitment email. This link took them to the consent form for the study. After consenting, participants began the survey. A summary of the survey questions and their types may be found in Table 5.

As the focus of the survey was on classroom teams, peer evaluation, and bias, students who began the study but had not participated in a classroom team (Table 6, Question 1) were redirected to the end of the survey and thanked for their interest. The study included 419 participants. Participants were allowed to skip questions with which they were uncomfortable.

Table 6. Questions in the student survey. * indicates demographics question

Question	Type
Have you ever worked in teams in a class or for a class project?	Yes/No
How much do you agree: I like working in teams on class projects	Likert (1-Strongly Disagree; 7-Strongly Agree)
How much do you agree: I avoid classes that involve teamwork?	Likert (1-Strongly Disagree; 7-Strongly Agree)
When working on a team in a class or for a class project, have you felt any biases from your teammates?	Yes/Maybe/No
Follow-up: Please describe the bias(es) you have felt from your teammates	Free response
Follow-up: How frequently have you felt bias from your teammates?	Likert (1-Never; 5-Always)
How much do you agree: When working on a team in a class or on a class project, I feel respected by my teammates	Likert (1-Strongly Disagree; 7-Strongly Agree)
How many classes have you taken that use peer assessment?	Multiple choice: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10+
How much do you agree: I feel that the peer assessments I have received were fair	Likert (1-Strongly Disagree; 7-Strongly Agree)
How much do you agree: I feel that the peer assessments I have given were fair	Likert (1-Strongly Disagree; 7-Strongly Agree)
Have you felt any biases in peer assessments you have received?	Yes/Maybe/No
Follow-up: Please describe the bias(es) you have felt in the peer evaluations you have received	Likert (1-Never; 5-Always)
In what ways, if any, do you think bias could affect the peer assessments you receive from classmates?	Free response
How much do you agree: I feel that the peer assessments I have given my classmates have been unbiased	Likert (1-Strongly Disagree; 7-Strongly Agree)

Table 6 Continued	
Question	**Type**
In what ways, if any, do you think bias could affect the peer assessments you give classmates?	Free response
*What is your gender?	Multiple choice: Woman, man, other, prefer not to say
*How would you describe yourself?	Multiple choice **(select all that apply):** White, Black or African American, Native American or Alaska Native, Asian, Native Hawaiian or Pacific Islander, Hispanic or Latinx, Other, Prefer not to answer
*Is English the first language you learned to speak?	Yes/no
**Follow-up:* How many years have you been speaking English?	Free numeric response
*How comfortable are you with speaking English?	Likert (1-Extremely Uncomfortable; 7-Extremely Comfortable)
*Which college are you affiliated with?	Agriculture and Life Sciences, Business, Design, Engineering, Graduate, Human Sciences, Liberal Arts and Sciences, Veterinary Medicine
*What is your major?	Free response
*What class level are you?	Multiple choice: Freshman, Sophomore, Junior, Senior, Graduate
*Are you an international student?	Yes/No

Data Analysis

Qualitative survey data was analyzed for overall counts and themes. Survey items which utilized Likert scales were analyzed using a t-Test for comparison among the demographic variables gender, race, international student status, and English language learner status. A 1x5 ANOVA was used to analyze the effect of class level. A significance level of $\alpha = .05$ was used throughout. Cohen's d was used to determine effect size (Cohen, 1988). The variable "gender" was simplified into two categories (Men and Women) due to lack of participants in the "other" and "prefer not to say" categories. The variable "race" was simplified into two categories: Students of Color (POC) and White.

Results

Means and standard deviations for each Likert survey question are reported in Table 7.

Table 7. Overall means and standard deviations of survey results, by question.

Question	Mean (SD)	N
Liking working in teams	4.1 (1.7)	342
Avoiding classes that involve teamwork	3.2(1.5)	342
Feeling respected by teammates	5.1 (1.2)	342
Classmates want to be on a team with me	5.0(1.3)	342
Peer assessments received were fair	5.5 (1.2)	342
Peer assessments given were fair	5.9(1.1)	342

Differences by Gender

There was a significant difference for liking working on teams, $t(328) = -3.02$, $p < .001$, $d = 0.36$. Women ($M = 3.9$, $SD = 1.7$, $N = 203$) liked teaming significantly less than Men ($M = 4.5$, $SD = 1.6$, $N = 133$). There was a significant difference for feeling respected by teammates, $t(328) = -3.19$, $p < .001$, $d = 0.35$. Women ($M = 5.1$, $SD = 1.2$, $N = 203$) felt significantly less respected than Men ($M = 5.5$, $SD = 1.1$, $N = 133$). There was a significant difference for perceiving peer assessments received as fair, $t(293) = -2.58$, $p = .005$, $d = 0.28$. Women ($M = 5.4$, $SD = 1.2$, $N = 181$) reported significantly lower fairness than Men ($M = 5.7$, $SD = 1.1$, $N = 120$). There were no significant differences by gender for avoiding classes that involve teamwork (p=.28), perceptions of teammates wanting to be on a team with the participant (p=.19), or feelings that the peer assessments the participant has given were fair (p=.19).

Difference by Race

There was a significant difference in whether students reported that they liked working on teams, $t(328) = -2.24$, $p = .013$, $d = 1.2$. White students ($M = 4.0$, $SD = 1.7$, $N = 286$) liked teaming significantly less than POC students ($M = 4.7$, $SD = 1.5$, $N = 42$). There were no significant differences by race for feeling respected by teammates ($p = .25$), avoiding classes

involving teamwork (*p = .18)*, fairness of peer assessments received (*p = .18)*, perceptions of teammates wanting to be on a team with the participant, (*p = .34)* or feelings that the peer assessments the participant has given were fair (*p = .16*)

Difference by English Speaker Status, International Student Status and Class Level

There was a significant difference for feeling the peer assessments received were fair, $t(301) = 1.8$, $p = .04$, $d = .52$. Students whose first language was not English ($M = 4.8$, $SD = 1.9$, $N = 19$) found the assessments less fair than native English speakers ($M = 5.6$, $SD = 1.1$, $N = 282$). There were no significant differences by international student status for liking working on teams (*p = .41)*, feeling respected by teammates (*p = .34)*, avoiding classes involving teamwork (*p = .30)*, perceptions of teammates wanting to be on a team with the participant (*p = .17)*, or feelings that the peer assessments the participant has given were fair (*p = .22)*. There were no significant differences for any of the items (liking working on teams (*p = .29)*, feeling respected by teammates (*p = .51)*, avoiding classes involving teamwork (*p = .44)*, perceptions of fairness (*p = .28)*, perceptions of teammates wanting to be on a team with the participant (*p = .33)*, feelings that the peer assessments the participant has given were fair) (*p = .25)* by international student status. There were no significant differences for any of the following items: liking working on teams (*p = .37)*, feeling respected by teammates (*p = .38)*, avoiding classes involving teamwork (*p = .19)*, perceptions of fairness (*p = .30)*, perceptions of teammates wanting to be on a team with the participant (*p = .25)*, feelings that the peer assessments the participant has given were fair (*p = .47)* by class level.

Occurrence of Bias

When reporting results, unless otherwise noted, “perceived bias” indicates the participant perceiving bias against themselves. Ninety-three participants (27.3%) reported they had felt bias

from their teammates while 63 (18.5%) might have felt bias and 185 had not felt bias (54.2%; N = 341). The most commonly felt biases were due to gender (31 mentions), race (14 mentions), age (10 mentions), major (11 mentions), and interpersonal relationships (20 mentions). Other sources of bias reported with less frequency were due to cultural differences, academic standing, and identification as LGBTQIA+. Of those students who had or might have experienced bias, the mean frequency of experiencing bias was 2.5 (SD = 0.8, N = 156).

Bias in Peer Assessments

Thirty-one participants (9.2%) reported they had felt bias in the peer assessments they had received while 31 (9.2%) might have felt bias and 275 had not felt bias (81.6%, N = 337). The most commonly felt biases were due to gender (13 mentions), personality (8 mentions), interpersonal relationships (10 mentions), age (5 mentions), "gaming the system"/obligation (11 mentions), and potential retaliation (5 mentions). Of those students who had or might have experienced bias in their peer assessments, the mean frequency of experiencing bias was 2.83 (SD = .89, N = 62).

How Bias Could Affect Peer Assessments Received

All participants were asked for examples of the ways in which bias *could* affect the peer assessments they *receive.* Seven participants chose to reflect on how they felt unlikely to be the recipient of a biased peer assessment. For example: "*None, really. To be frank, I'm a straight white dude so not really subject to a lot of the race/sex-based biases that impact academia and many industries."* For the rest of the participants, friendship status (42 mentions), "gaming the system"/obligation (34 mentions), and personality (28 mentions) were the most commonly mentioned ways in which they could be the target of bias. Other items mentioned were gender, race, age, ability, and perceived fluency in the language of class instruction.

How Bias Could Affect Peer Assessments Given

All participants in the study were asked for examples of the ways in which bias *could* affect the assessments they *give* to their peers. Many participants (14) responded that they try to be impartial or that bias would not affect the peer assessments they give. For the rest of the participants, friendship status (41 mentions), "gaming the system"/obligation (33 mentions), and personality (29 mentions) were the most commonly mentioned ways in which they could give biased assessments. Other items mentioned were gender, age, perceived ability, and language.

Method and Results: Thinkspace Peer Assessment Data Analysis

An Institutional Review Board (IRB) approved analysis of peer assessment data was conducted using peer assessment data from Iowa State University within the Thinkspace learning tool (Appendix F). This data was linked to demographic data provided by the Registrar using a key and analyzed for evidence of bias.

Objective

In the first two studies described in this chapter, the issue of peer assessment bias was explored from both the student and instructor perspective through subjective survey data. In this study, a large body of peer assessment data from the Thinkspace peer assessment tool will be analyzed for bias.

Participants

Data from 3,885 students was included in the analysis. The study included 1,972 females and 1,913 males. Table 8 gives the breakdown of participants by race or ethnicity. It should be noted, however, that the Office of the Registrar codes international students as "International" as opposed to racial/ethnic categories such as "white" or "Asian". Students coded as "International" could be of any race or ethnicity.

Table 8. Race or ethnicity of students in the Thinkspace dataset

Race or Ethnicity	Count
White (not Hispanic)	2,964
International	246
Prefer not to indicate	185
Hispanic (Spanish American)	181
Asian	120
Black (not Hispanic)	98
Multiracial	77
American Indian or Alaskan Native	8
Native Hawaiian or Pacific Islander	3

Five-hundred ten students were first generation college students, while 3,374 were not. English was the first language of 3,576 students, while 309 students initially spoke another language. Students ranged in age (in 2020) from 20 to 62 years. The mean and median ages were both 25 years. Two-hundred forty-six students were non-United States natives (international students), while 3,639 students were United States natives. Federal Pell Grants are grants awarded to undergraduate students who display exceptional financial need and have not yet earned a bachelor's, graduate, or professional degree (Federal Student Aid, 2020). Pell grants were received by 760 students and not received by 3,125 students. A breakdown of the students in the study by academic college is given in Table 9.

Table 9. Academic college affiliation of the students in the Thinkspace dataset

College	Count
Human Sciences	899
Engineering	889
Liberal Arts and Sciences	851
Agriculture and Life Sciences	769
Business	300
Design	140
Interdisciplinary	18
Veterinary Medicine	6

Peer Assessment Procedure

Thinkspace supports two main methods of peer evaluation: Categories and Balance (Michaelsen). In the categories method, students are given a set of categories or questions to evaluate their peers by. These categories or questions may vary among instructors. The balance method was developed by Michaelsen et al. (1997) and gives students a set number of points to distribute among their peers. In the balance method of peer assessment, each student is given a number of points to distribute among team members. The most common method is to set each assessed student in a team to a "worth" of 10 points. For example, if a team had 5 members, and each member was "worth" 10 points, the total number of points a student would have available to split among their 4 team members would be 40 (the reviewing student does not review themselves.) The number of points available to distribute may vary with team size and instructor preference, but the core of the method remains the same across all instructors. Figure 2 shows the "student view" of a balance method peer assessment in Thinkspace.

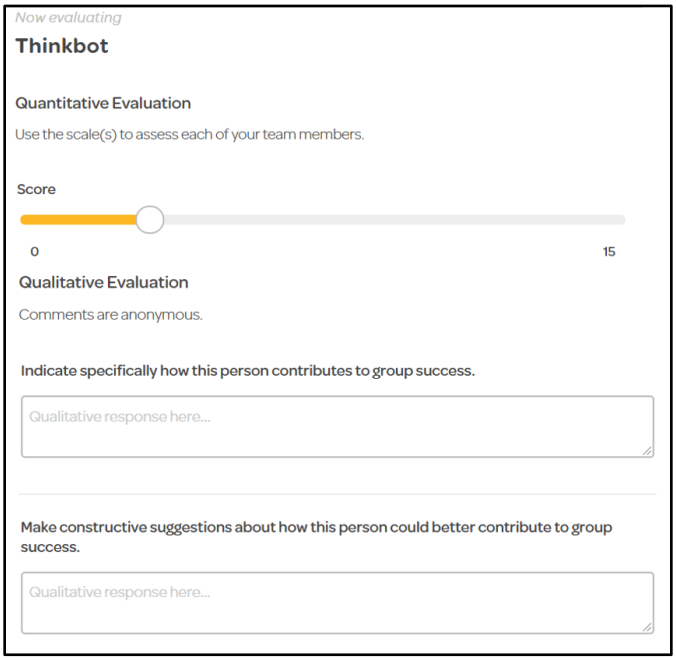

Figure 2. Student view of a Michaelsen (balance) peer assessment in Thinkspace

Data Analysis Procedure

The data used for this analysis was provided by Thinkspace and contained 24,180 rows of individual balance method peer assessments completed between the Spring 2013 semester and the Fall 2018 semester. Only the balance method data was used in the analysis as this method is standardized across classes, while the categories method is not. Before linking the peer assessment data to demographic data, it was "cleaned" using the following strategies, which resulted in the removal of 1,123 rows.

1. Thinkspace allows a "sandbox" mode for instructors to explore the environment. These sandboxes were included in the provided data, but were removed prior to analysis. Sandboxes were recognized by their title (often involving the word "sandbox"), or the title of the peer assessment itself (again often involving the word "sandbox" or "test").
2. Some peer assessments were set up within Thinkspace but never used. These items were recognized by their all blank score and qualitative feedback sections.
3. The Thinkspace tool has also been used in educational workshops and demos. These items were recognized by titles such as "demo" or "workshop".

The peer assessment data was then linked to demographic data using the reviewer and reviewee ID codes. Each peer assessment rating contained a code for the individual who gave the rating (the reviewer) and the individual who received the rating (the reviewee). On the left side of the table, the data was linked using the reviewer code to the reviewer demographics. Similarly, on the right side of the table, the data was linked using the reviewee code to the reviewee demographics (Figure 5).

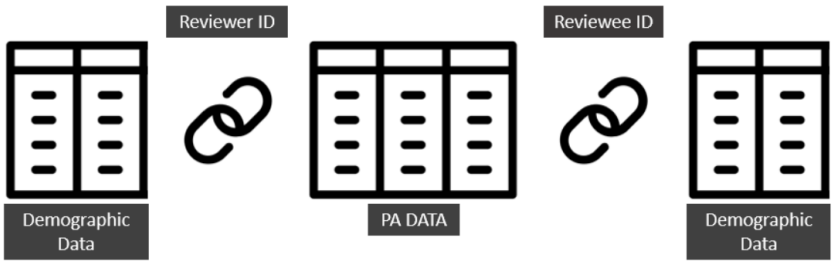

Figure 3. Procedure for linking demographic data with peer assessment data

After linking, the data was cleaned once again. While all rows of peer assessment data contained both a reviewer and reviewee code, the data provided by the registrar did not include information for every ID code. If no demographic information was present, the row was removed as the analysis to be performed relied on knowledge of demographics. This resulted in the removal of 547 rows, for a total of 22,510 remaining rows of data – each containing one peer assessment rating.

However, some instructors in the dataset chose to use a worth value per teammate other than 10. For these classes, the peer assessment data had to be standardized to be on the same scale as the rest of the data. For example, suppose Class A used a worth of 20 points per team member. In the dataset, all ratings associated with Class A would then be divided by 2 to put them on the same scale as the rest of the ratings. The final dataset included 22,510 ratings given by 3885 students in 115 classes. A summary of the items included in the dataset is given in Table 10.

Table 10. Items included in the final dataset after linking between registrar data and peer assessment data

DEMOGRAPHICS		THINKSPACE
One-Time	**Per Semester F2013-S2018**	----------
Sex	Level (Undergrad/Graduate)	Assignment ID
Ethnicity	Classification (Freshman, Sophomore, Junior, Senior, Graduate)	Phase ID/Phase Title
Interim ZIP	Major	Contributions (Qualitative)
First Gen Student (Y/N)	College	Constructive (Qualitative)
Language	# of Midterms (Grades C- and below)	Quantitative Score
High School	Semester GPA	Space ID/Space Title
Birth M/Y	Cumulative GPA	
Country of Origin		
Pell Grant Status (Y/N)		

Peer assessment scores were analyzed using a multiway ANOVA with student's T-Test or Tukey post-hoc analysis for significant results. Specific analyses included:

- effect of sex on peer assessment score (given and received)
- effect of sex on GPA
- effect of ethnicity on peer assessment score (given and received)
- effect of international student status on peer assessment score (given and received)
- effect of English speaker status on peer assessment score (given and received)
- effect of Pell grant status on peer assessment score (given and received)

A significance level of $\alpha = .05$ was used throughout. Effect size was determined using Cohen's d (Cohen, 1988).

Results

Effect of sex on peer assessment score

There was a significant main effect of Reviewee Sex on peer assessment score received, $F(1, 22{,}509) = 23.7$, $p < .001$, $d = 0.38$). The mean score *received* by males ($M = 10.1$, $SD = 0.8$) was significantly higher than the mean score received by females ($M = 9.8$, $SD = 0.8$). There was a significant main effect of Reviewer Sex on peer assessment score given, $F(1, 22{,}509) =$

228.3, $p < .001$, $d = 0.47$). The mean score *given* by females ($M = 10.2$, $SD = 0.9$) was significantly higher than the mean score given by males ($M = 9.7$, $SD = 0.9$). There was also a significant interaction between the gender of the reviewer and the gender of the reviewee, $F(1, 22{,}509) = 16.3$, $p < .001$). The mean score given by females rating males was significantly higher than the mean score given by males rating males ($p < .001$, $d = 0.18$) and males rating females ($p < .001$, $d = 0.36$). There was no significant difference between mean score given by females rating males ($p = .20$) and females to females ($p = .23$). The mean score given by males rating males was significantly lower than the mean score given by females rating males ($p < .001$, $d = 0.18$), and females rating females ($p < .001$, $d = 0.24$). The mean score given by males rating males was significantly higher than males rating females ($p < .001$, $d = 0.18$). These results for the interaction of reviewer and reviewer gender are summarized in Table 11.

Table 11. Means and standard deviations by sex of reviewer and reviewee. Levels not connected by the same letter are significantly different.

Reviewer (given)	Reviewee (received)	Letter Report	Mean	SD
F	M	A	10.2	1.6
F	F	A	10.2	1.7
M	M	B	9.9	1.8
M	F	C	9.6	1.7

Effect of Ethnicity on Peer Assessment Score

The analysis showed a significant main effect of reviewee ethnicity on peer assessment score received $F(2, 22{,}507) = 36.2$, $p = <.001$. The mean score received by white students ($M = 10.1$, $SD = 1.9$) was significantly higher than the mean score received by students of color ($M = 9.7$, $SD = 2.0$, $p < .001$, $d = 0.22$) and international students ($M = 9.5$, $SD = 2.6$, $p < .001$, $d = 0.24$). The mean score given by white students ($M = 9.6$, $SD = 2.0$) was significantly lower than the mean score given by international students ($M = 9.9$, $SD = 2.0$, $p < .001$, $d = 0.19$). There were no

significant differences in the scores given by students of color and international or white students. Additionally, there were no significant differences in the scores received by students of color and international students. There was also a significant interaction between the ethnicity of the reviewer and the ethnicity of the reviewee. These results for the interaction of reviewer (given) and reviewee (received) ethnicity are summarized in Table 12 and Table 13.

Table 12. Means and standard deviations by ethnicity of reviewer and reviewee. Levels not connected by the same letter are significantly different

Reviewer (given)	Reviewee (received)	Letter Report	Mean	SD
International	White	A	10.2	2.0
Students of color	White	A	10.1	1.8
White	White	A	10.1	1.9
International	Students of Color	A B	9.9	2.1
International	International	A B	9.7	2.2
White	Students of color	B	9.6	2.0
Students of color	International	B C	9.6	1.5
Students of color	Students of color	B C	9.6	2.1
White	International	C	9.2	1.7

Table 13. Statistics for the interaction of reviewer and reviewee ethnicity. Means for the levels in the first column are significantly higher than means for the second column.

Level (reviewer, reviewee)	Level (reviewer, reviewee)	*p*	*d*
International, White	White, International	<.001	0.53
Students of color, White	White, International	<.001	0.52
White, White	White, International	<.001	0.49
International, Students of color	White, International	.002	0.36
International, White	Students of color, Students of color	<001	0.34
International, White	Students of color, International	.022	0.35
Students of color, White	Students of color, Students of color	<.001	0.32
International, White	White, Students of color	<.001	0.35
White, White	Students of color, Students of color	<.001	0.26
International, International	White, International	.016	0.29

Table 13 Continued

Level (reviewer, reviewee)	**Level (reviewer, reviewee)**	***p***	***d***
Students of color, White	Students of color, International	.038	0.28
Students of color, White	White, Students of color	<.001	0.34
White, White	Students of color, International	.039	0.27
White, White	White, Students of color	<.001	0.31
White, Students of color	White, International	<.001	0.35

Effect of International Student Status on Peer Assessment Score

The analysis showed a significant main effect of reviewee international student status on peer assessment score received $F(1, 22{,}506) = 55.3$, $p = <.001$. The mean score received by US students ($M = 10.1$, $SD = 1.9$) was significantly higher than the mean score received by International students ($M = 9.5$, $SD = 2.6$, $p <.001$, $d = 0.31$). The mean score given by US students ($M = 9.7$, $SD = 2.0$) was significantly lower than the mean score given by International students ($M = 9.9$, $SD = 2.1$, $p <.001$, $d = 0.18$). There was also a significant interaction between the country of origin of the reviewer and the country of origin of the reviewee. The mean score given by US students rating international students was significantly lower than the mean score given by International students rating US students ($p <.001$, $d =0.47$), The mean score given by US students rating US students ($p<.001$, $d = 0.40$), was significantly higher than International students rating International students ($p = .007$, $d = 0.24$). There were no significant differences between mean score given by International Students rating US Students and US Students rating US Students ($p = .33$). There were no significant differences between mean score given by International students rating US Students and International Students rating International Students ($p = .20$). These results for the interaction of reviewer and reviewee international student status are summarized in Table 14.

Table 14. Means and standard deviations by international student status of reviewer and reviewee. Levels not connected by the same letter are significantly different

Reviewer (given)	Reviewee (received)	Letter Report	Mean	SD
International	US	A	10.1	1.9
US	US	A	10.0	2.0
International	International	A	9.7	2.0
US	International	B	9.3	2.2

Effect of English Speaker Status on Peer Assessment Score

The analysis showed a significant main effect of reviewee English speaker status on peer assessment score received $F(1, 22{,}509) = 65.4$, $p = <.001$. The mean score received by native English speakers ($M = 10.1$, $SD = 1.9$) was significantly higher than the mean score received by non-native English speakers ($M = 9.5$, $SD = 2.2$, $p <.001$, $d = 0.28$). The mean score given by native English speakers ($M = 9.7$, $SD = 2.1$) was significantly lower than the mean score given by non-native English speakers ($M = 9.9$, $SD = 1.8$, $p<.001$, $d = 0.13$). There was also a significant interaction between the country of origin of the reviewer and the country of origin of the reviewee. The mean score given by native English speakers to non-native English speaking students ($M = 9.3$, $SD = 2.1$) was significantly lower than the mean score given by non-native English speakers rating English speakers ($M = 10.1$, $SD = 2.0$, $p < .001$, $d = 0.41$), native English speakers rating native English speakers ($M = 10.0$, $SD = 1.9$, $p <.001$, $d = 0.38$), and non-native English speakers rating non-native English speakers ($M = 9.7$, $SD = 2.0$, $p <.032$, $d = 0.21$). There were no significant differences between mean score given and received in any other pairs. These results for the interaction of reviewer and reviewee language are summarized in Table 15.

Table 15. Means and standard deviations by English speaker status of reviewer and reviewee. Levels not connected by the same letter are significantly different

Reviewer (given)	Reviewee (received)	Letter Report	Mean	SD
Non-native	Native	A	10.1	2.0
Native	Native	A	10.0	1.9
Non-native	Non-native	A	9.7	2.2
Native	Non-native	B	9.3	2.1

Effect of Pell Grant Status on Peer Assessment Score

The analysis showed a significant main effect of reviewee Pell grant status on peer assessment score received $F(1, 22{,}509) = 22.4$, $p = <.001$. The mean score received by non-Pell grant recipients ($M = 10.1$, $SD = 2.0$) was significantly higher than the mean score received by Pell grant recipients ($M = 9.8$, $SD = 2.4$, $p < .001$, $d = 0.15$). There were no significant differences for the scores given by Pell grant recipients and non-Pell grant recipients. There was also a significant interaction between the Pell grant status of the reviewer and the Pell grant status of the reviewee. The mean score given by non-recipients to recipients ($M = 10.1$, $SD = 2.1$) was significantly higher than the mean score given by recipients to non-recipients ($M = 9.8$, $SD = 1.9$, $p <.001$, $d = 0.14$) and recipients to recipients ($M = 9.7$, $SD = 2.3$, $p <.001$, $d = 0.16$). The mean score given by non-recipients to non-recipients ($M = 10.0$, $SD = 2.1$) was significantly higher than the score given by recipients to non-recipients ($p <.001$, $d = .14$) and recipients to recipients ($p < .001$, $d = 0.16$). There were no significant differences between mean score given and received in any other pairs. These results for the interaction of reviewer and reviewee Pell grant status are summarized in Table 16.

Table 16. Means and standard deviations by Pell grant status of reviewer and reviewee. Levels not connected by the same letter are significantly different

Reviewer (given)	Reviewee (received)	Letter Report	Mean	SD
Non-recipient	Recipient	A	10.1	2.1
Non-recipient	Non-recipient	A	10.0	2.1
Recipient	Non-recipient	B	9.8	1.9
Recipient	Recipient	B	9.7	2.3

Effect of Demographics on GPA

To determine if differences in GPA could explain the differences in peer assessment score by sex, international student status, native language, and Pell Grant status, we conducted an analysis of the effect of these demographic variables on GPA. There was a significant effect of sex on GPA $F(1, 22{,}509) = 629.4$, $p < .001$, $d = 0.33$) where the mean GPA for female students ($M = 3.14$, $SD = .59$) was significantly higher than the mean GPA for male students ($M = 2.93$, $SD = .67$). There was a significant effect of international student status on GPA ($F(1, 22{,}509) = 227.5$, $p < .001$, $d = 0.38$) where the mean GPA for international students ($M = 3.28$, $SD = .68$) was significantly higher than the mean GPA for US students ($M = 3.03$, $SD = .63$). There was a significant effect of Pell Grant status on GPA ($F(1, 22{,}509) = 376.3$, $p < .001$, $d = 0.32$) where the mean GPA for non-Pell Grant recipients ($M = 3.08$, $SD = .63$) was significantly higher than the mean GPA for Pell Grant recipients ($M = 2.88$, $SD = .61$). There was a significant effect of language on GPA ($F(1, 22{,}509) = 21.6$, $p < .001$, $d = 0.14$) where the mean GPA for students whose first language was not English ($M = 3.12$, $SD = .69$) was significantly higher than the mean GPA for students whose first language was English ($M = 3.03$, $SD = .63$).

The effect of race on GPA was analyzed using ANOVA. There was a significant main effect of race on GPA $F(2, 22{,}508) = 258.3$, $p < .001$). The mean GPA for international students ($M = 3.28$, $SD = .67$) was significantly higher than the mean GPA for white students ($M = 3.05$, $SD = .62$, $p < .001$, $d = 0.36$) and students of color ($M = 2.84$, $SD = .64$, $p < .001$, $d = 0.67$). The mean

GPA for white students (M =3.05, SD =.62) was significantly higher than the mean GPA for students of color (M =2.84, SD =.64, p <.001, d = 0.33). These results are summarized in Table 17.

Table 17. Means and standard deviations of GPA by race. Levels not connected by the same letter are significantly different

Race	Letter Report	Mean	SD
International	A	3.28	.67
White	B	3.05	.62
Students of Color	C	2.84	.64

Discussion

The three studies reported above established evidence of bias in peer assessment. In the student survey, participants noted bias in the assessments they had received and reflected on how the peer assessments they gave could be biased. Similarly, in the instructor survey, participants once again noted bias in their classrooms and assessments. Finally, the evidence of bias is shown in peer assessment scores themselves, which cannot be fully explained by student achievement (e.g. GPA).

Perception of bias in the classroom

The first study indicated that a large portion of instructors (47%) have perceived bias in their peer evaluations. These results are even higher than those reported by Vogel et al. in 2009 where 27% of professors and 25% of graduate instructors noticed bias in their classrooms. The types of biases observed in both studies are also similar with race, sex, and country of origin being represented. These perceptions are given further credibility by the results of study three which show that sex, race, and country of origin can have a negative effect on peer evaluation scores. While many of the biases described by the instructors in this study were what often comes to mind when reading the word “bias” (e.g. racism, sexism), others were specific to

collaborative learning environments and peer assessment, such as a general unwillingness to give negative ratings or feedback. This unwillingness to be critical of peers has been observed in other studies of peer assessment (e.g. Topping, 2009; Arnold, et al., 2005).

Mitigation of Bias

While none of the studies focused specifically on bias mitigation, instructor responses to a question on their mitigation efforts provide direction for future work in this area. Over half the instructor participants had taken action to mitigate bias in their peer assessments. These actions ranged from using the built-in calibration function within their peer assessment platforms to feedback on what constitutes a constructive comment prior to administering the assessments. Some instructors chose to mitigate bias through team composition by intentionally diversifying teams in terms of gender, race, and background. One instructor mentioned taking this a step further by diversifying teams while ensuring that no team member was the "only" person from a group on the team (e.g. the only woman or only person of color). Both of these strategies take advantage of one of the greatest predictors of bias: contact with marginalized groups. Positive contact with members of marginalized groups has been shown to reduce bias even when the contact is not sustained (such as a guest speaker in a classroom) and reduce bias most effectively when sustained over time (such as through a permanent learning team) (Burke, Dovidio, Przedworski, Hardeman, Perry, Phelan & Nelson, 2015; Herek & Capitanio, 1996; Kelley, Chou, Dibble, Robertson, 2008).

Bias of personality

Another bias to note is often described as "personality". While there is no data on the personality types associated with the biases instructors noticed, clues can be found in the written descriptions of peer assessment bias that were provided. Of those which were personality related,

positive bias was mentioned in conjunction with descriptors such as “outgoing”. This type of description is often associated with the expressive social style, which has been shown to receive higher peer evaluation scores in studies of oral presentations (Valencia, Carrillo, & Benítez, 2012). Conversely, negative bias was often mentioned in conjunction with descriptors such as “quiet” or “decisive”. This type of description follows the “analytical” or “driver” social styles, which have been associated with lower peer assessment ratings (Valencia, Carrillo, & Benítez, 2012). These differences in peer assessment ratings due to social style have been able to be mitigated through training (May, 2008). Therefore, inclusion of social style in future trainings could be beneficial.

Student Perceptions of Peer Assessment

In the second study, women reported significantly lower levels of enjoyment of working in teams, lower levels of respect from teammates, and lower perceptions of fairness in their peer assessments than men. Similarly, 66% of the students who had experienced bias in a classroom team and 60% of the students who had experienced bias in their peer assessments were women. These findings for both attitude toward classroom teams and experience with peer assessment are consistent with work that has shown that male students report more positive attitudes about peer assessment than female students (Wen, & Tsai, 2006; Topping, 2010). The results of the third study demonstrate that these lower perceptions of fairness reflect the reality of the peer assessment scores women earn.

Student participants also made many of the same observations as instructor participants in terms of bias in peer assessment due to personality and a reluctance to give negative feedback. Descriptions of “personality” bias once again follow the general descriptions of social style with those who exhibit the expressive style often being portrayed as receiving higher ratings.

However, many students made specific mention of bias due to friendship, which is a commonly noted potential source of biased peer assessments (Brown & Knight, 1994; Langan, Wheater, Shaw, Haines, Cullen, Boyle, Penney, Oldekop, Ashcroft & Lockey, 2005).

In addition to biases they observed, students were asked for ways in which bias *could* affect the assessments they give and receive. Respondents often focused responses on the potential for personality and friendship to influence their ratings. However, some students noted categorical biases (e.g. gender, language) that could influence how they rate and are rated. This indicates that these students are aware of the potential for bias which is an important first step in mitigation (Devine, et al., 2012). When reporting on biases they had perceived against themselves as well as the ways in which peer assessments *could* be biased, students mentioned the interconnected issues of retaliation and a general unwillingness to give negative feedback. The attention paid to these issues by participants shows that some students are coming into the peer assessment process already affected by prior experiences (e.g. experiencing retaliation or knowing that students have used peer assessments as retaliation). These prior experiences may unwittingly lead students to rate their peers less honestly than they otherwise would.

Fairness of Rating by Gender

In Study 3, female raters received lower scores overall, yet gave higher scores overall, which is similar to findings by Bryan and colleagues (2005) and Sherrard and colleagues (1994). This is in contrast to May and colleagues (2008) who found that men received lower ratings overall. The finding for men receiving lower overall ratings is often explained by women earning higher GPAs (e.g. May and Gueldenzoph, 2006). This was also true in the current study, but because females received the lower peer assessment scores, GPA does not fully explain this discrepancy. Unlike the studies referenced here (e.g. Espey, 2021; May et al., 2008; May and

Gueldenzoph, 2006), however, the current work analyzed a large body of data outside of the realm of a single classroom or department. Further, female raters showed no significant differences in the ratings they gave to males or other females. This finding could suggest the female raters were fairer in terms of gender.

Fairness of Rating by International Status and Race

While the effect of gender on peer assessment scores has received considerable amount of study, the effect of international student status has received much less. International student enrollment in the United States has greatly increased in the past twenty years, however, these students are often the targets of nativism, racism, and other forms of discrimination (Yao, George Mwangi, & Malaney Brown, 2019; Lee & Rice, 2007). In the analysis of peer assessment scores, US students gave international students lower peer assessment scores than they gave their US-based peers. When race was added to the analysis, the outcomes remained the same. White domestic students received higher scores than international students of any race. Conversely, white domestic students rate international students lower than other domestic white students. In both instances, international students had higher GPAs than white students and US-based students, so their lower peer assessment scores are less likely to be explained differences in achievement.

In the analysis of peer assessment scores by race, white students received higher scores than students of color. Conversely, white students rated students of color lower than other white students. These results showing that the contributions of students of color are less valued are similar to those found by Wayland, et al., 2014. Unlike the comparisons of peer assessment scores with GPA made for gender, students of color in this study had both lower peer assessment scores and lower GPAs. However, there are numerous interacting factors specific to the

experiences of students of color that may affect GPA. Students of color, specifically Black, Latinx, and Native American students historically earn GPAs that are lower than their white counterparts (Fletcher, & Tienda, 2010; Woo, Green, & Matthews, 2013). Students from these groups are also disproportionally working learners from low-income backgrounds (Carnevale & Smith, 2018). This suggests that, since students of color are more likely to come from lower-income backgrounds and work longer hours while in college, their GPAs suffer. Quality of high school attended is also a contributing factor to the disparities in GPAs earned by students of color and white students (Fletcher, & Tienda, 2010). Minority students often attend high schools with lower instructional quality than white students, which has been shown to be a factor in differences in college achievement (Fletcher, & Tienda, 2010). As a result, the lower peer assessment scores received by students of color cannot be fully separated from the interconnected factors of bias, student team performance, and GPA.

For women and international students, peer assessment scores and GPA move in opposite directions. Conversely, for students of color, peer assessment scores and GPA move in similar directions. When interpreting the results of these studies, it should be noted that GPA is not suggested to be analogous with, or a direct predictor of, peer assessment score. Some work has shown that GPA and peer assessment score are correlated overall with students earning higher GPAs also earning higher peer assessment scores, however the scope of this work is limited (Al Mortadi, Al-Houry, Alzoubi, & Khabour, 2020). Limitations in applying the results of Al Mortadi, et al., 2020 to the present study include all participants being students in a dental graduate program, overall small sample size (130 students). The present analysis of GPA and peer assessment score was conducted because GPA has been suggested as a potential factor in differing peer assessment scores (e.g. May and Gueldenzoph, 2006). Nevertheless, it is possible

that other factors (e.g. student team performance) are effecting peer assessment scores more than GPA or bias.

Limitations

All three studies were limited by the demographics of the institution in which they were deployed. As most participants were white and from the United States, the results of these studies are not generalizable to all students or instructors, especially students of color and international students and their instructors. The analysis of the peer assessment in the third study is limited by the coding of student data. "International" is not a race or ethnicity, and the use of this code instead of the actual race or ethnicity of the international student makes it difficult to fully understand the issue of bias for students of color.

Contribution

This work contributes new information into the body of knowledge on peer assessment bias, specifically by targeting broad participant groups across classrooms and disciplines. The results from these studies are valuable themselves, but will also contribute to the design of training to increase the fairness of peer assessments by giving the designers insight into what types of bias have the greatest impact on students.

CHAPTER 5: DEVELOPMENT AND FORMATIVE EVLAUTION OF TRAINING TO IMPROVE PEER ASSESSMENT FAIRNESS

Research Objectives and Introduction

The preceding chapter assessed the problem of bias in peer assessment. In this chapter, an initial iteration of training to address this issue will be described and evaluated.

Training to reduce general incidence of implicit bias in the classroom has been developed and yielded promising results (e. g. Devine et al., 2012) and focuses on drawing the trainee's attention to bias and its consequences as well as providing them with strategies to change biased behavior. However, training to reduce bias in peer assessment has been less explored, and generally focused on use in a single classroom (e.g. May, 2008).

Training on appropriate peer assessment criteria (e. g Onyia and Allen, 2012) has been studied and provides guidance for assessing aspects of team performance. However, specific training on how to mitigate bias when rating peers, and how to recognize and reframe biased language in peer assessment comments, has received less attention.

This work combines work on bias reduction and appropriate assessment to create bias training specifically for classes using peer assessment. The specific objectives of this study were to design, implement, and evaluate such a training in university classrooms. The design and implementation phases drew upon existing bias reduction methodologies, principles of effective feedback, Team-Based Learning literature, and best practices for training design. The evaluation of the training focused on student attitudes and perceptions of fairness at the beginning and end of the semester, as well as incrementally with each peer assessment.

Peer Assessment Fairness Training Design

The training was developed for deployment in Qualtrics in order to facilitate ease of implementation by instructors and eliminate the need for use of a class period. The training was divided into two parts: Part One: Giving Feedback and Part Two: Reducing Bias. Part One focused on the general process of giving useful and appropriate feedback in a peer evaluation while Part Two focused on recognizing and mitigating potentially biased ratings and comments. The outline of the training is shown in Figure 4. After learning information via readings or videos, participants completed activities where this information was put to use. It has long been established that feedback is essential to effective learning (Bellon, Bellon, and Blank, 1991; Race, 2001; Yorke, 2002). Therefore, feedback was presented to students each time they submitted an answer to an activity.

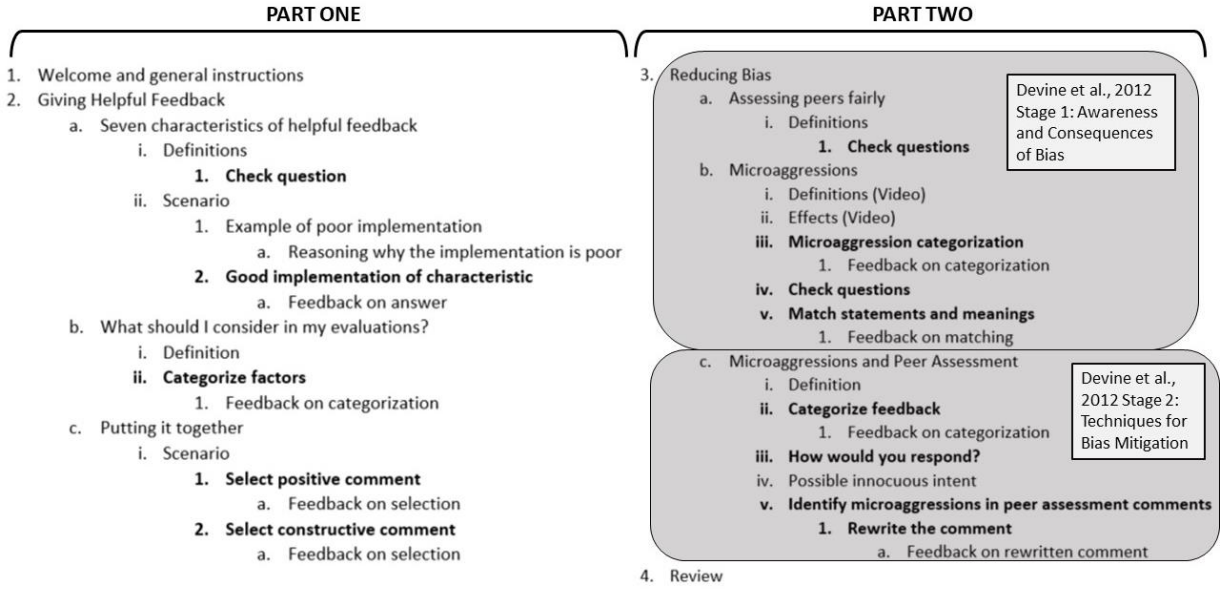

Figure 4. Outline of the training. Items in bold denote activities.

Part One: Seven Characteristics of Helpful Feedback

The material in the first part of the training drew heavily from classic Team Based Learning literature (e.g. Michaelson and Schultheiss, 1988). The first items covered were the seven characteristics of helpful feedback (Michaelson and Schultheiss, 1988) (Table 18).

Table 18. Definitions of the seven characteristics of helpful feedback (Michaelsen and Schultheiss, 1988)

Characteristic	Definition
Descriptive, not evaluative and is "owned" by the sender	Evaluative feedback expresses judgment of the receiver, or his or her actions. To assess peers well, we should objectively describe problems rather than speak in an evaluative manner. Evaluative assessment puts the listener on the defensive. "Owned" statements such as "I think..." or "I disagree..." describe the giver's position and invite comparison from the receiver.
Specific, not general	General statements of problems are often too large to be resolved and tend to oversimplify or misrepresent an issue. Specific feedback gives the receiver clear direction with little room for misinterpretation.
Honest and Sincere	It is helpful to provide direct, honest, and respectful feedback to your peer. Many people feel that they have to give a complement before they can say something constructive, but that can hide the important information they are trying to convey and prevent the recipient from hearing the message. Feedback often is most helpful when you get right to the point (NO extra words) and use every-day language (i.e., normal for you and appropriate to the setting).
Expressed in terms relevant to the needs of the receiver	Even though a behavior may be undesirable from your point of view, your feedback is likely to be ignored unless it is given in terms that are important to the recipient. For example, depending on their needs, telling a peer "I thought the way you treated Susan was unprofessional," might have a dramatically different effect than asking "Were you aware that Susan was so upset she was in tears?" Some might respond better to the first piece of feedback because they are concerned about maintaining professionalism, while others may respond better to the second piece of feedback because of Susan's reaction. When giving feedback, consider the group's norms and shared values.
Timely and in context	In general, the more immediate the feedback; the more helpful it will be. In part, this is because immediate feedback tends to be much more specific since the details of the situation are more apparent than they would be at any later point in time. In addition, delayed feedback often causes resentment because the recipient may feel that he or she could have minimized problems by making on-the-spot corrections if you had spoken up earlier.
Desired by the receiver	One of the most critical aspects of giving feedback is being able to tell when those who need it are ready to receive it. In part, this is because imposing feedback on someone who isn't ready for it is more likely to damage your relationship with the recipient than to provide him or her with helpful insights. Most people give both nonverbal and verbal cues about their willingness to receive feedback. For example, both body position and such things as attempting to redirect the conversation are warnings about giving negative feedback. Peer evaluations provide an appropriate time to give feedback, which makes it easier to receive that feedback and learn from it.

Table 18 Continued	
Characteristic	**Definition**
Usable and concerned with behavior over which the receiver has control	Feedback is useful only when it relates to something over which the person has control. Feedback is unhelpful when it is about personal attributes such as race, gender, age, physical size or even previous experience. The problem is that the person can't do anything about them even if they want to. As a result, giving feedback based on these kinds of qualities is not only unhelpful but is likely to cause resentment (or worse).

Each definition was accompanied by a scenario a student may encounter in a classroom peer assessment as well as a poor example of how the characteristic could be implemented in that scenario. The participant was then asked to rewrite the piece of feedback using a good implementation of the characteristic (Figure 5).

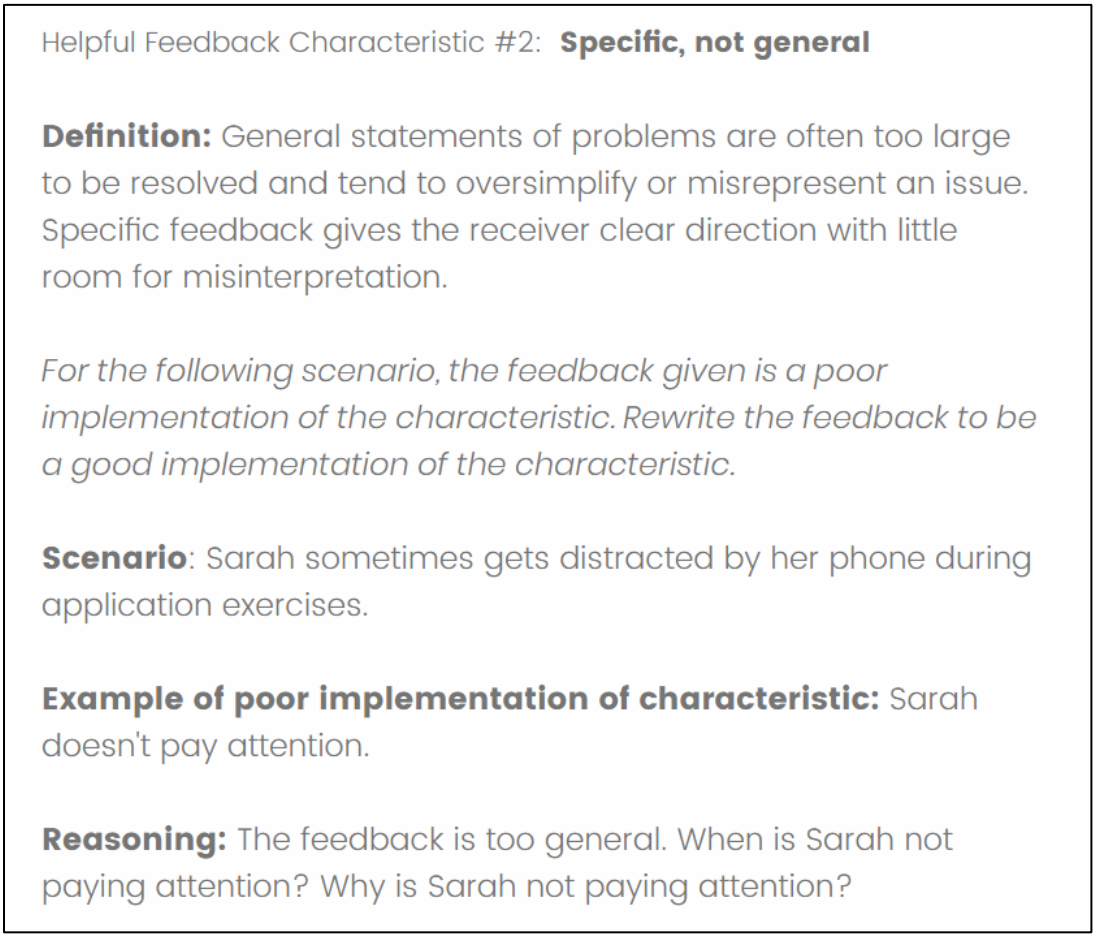

Figure 5. Example of a feedback characteristic, its poor implementation, and reasoning

After submitting their rewritten feedback, participants also received feedback on their answer (Figure 6).

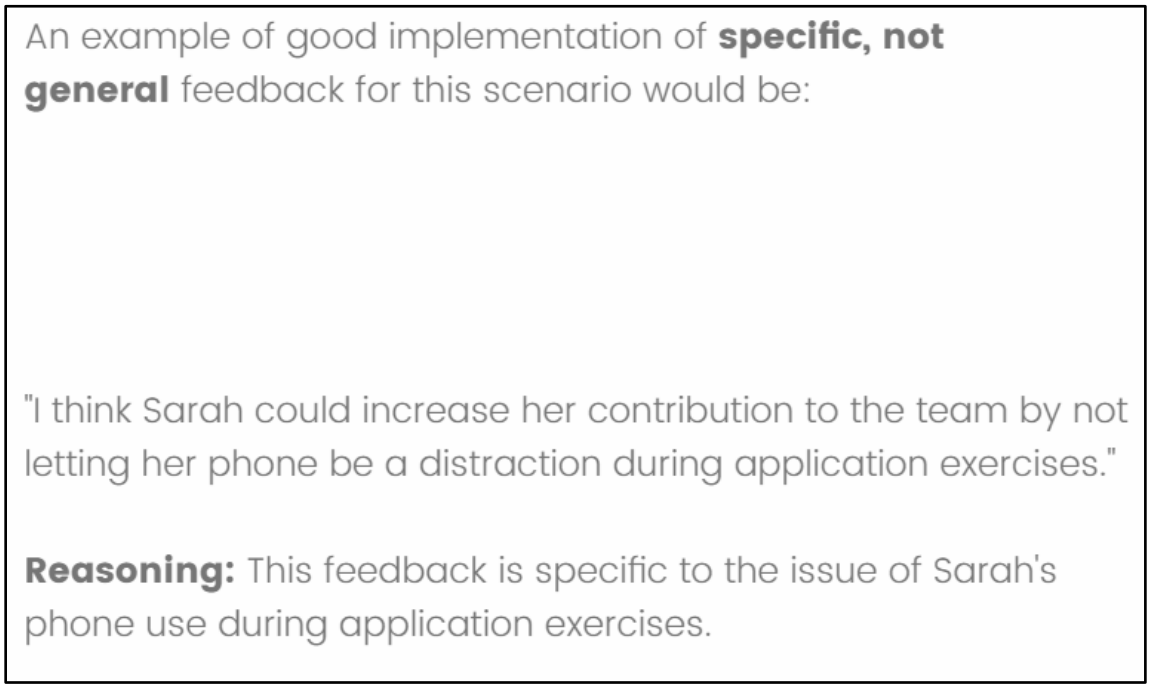

Figure 6. Feedback on answer with accompanying reasoning

Part One: What should I consider in my evaluations?

The next section of the training focused on what students should consider while evaluating their peers. The criteria presented were drawn from criteria by Okey Onyia and Stephanie Allen as part of their "Peer-Assessment Criteria Template (PACT) (Onyia & Allen, 2012) and input from instructors via the instructor survey described in Chapter 4. The PACT criteria were used as they demonstrated high student acceptance and comprehension (Onyia & Allen, 2012). The text of the section on criteria is presented in Figure 7.

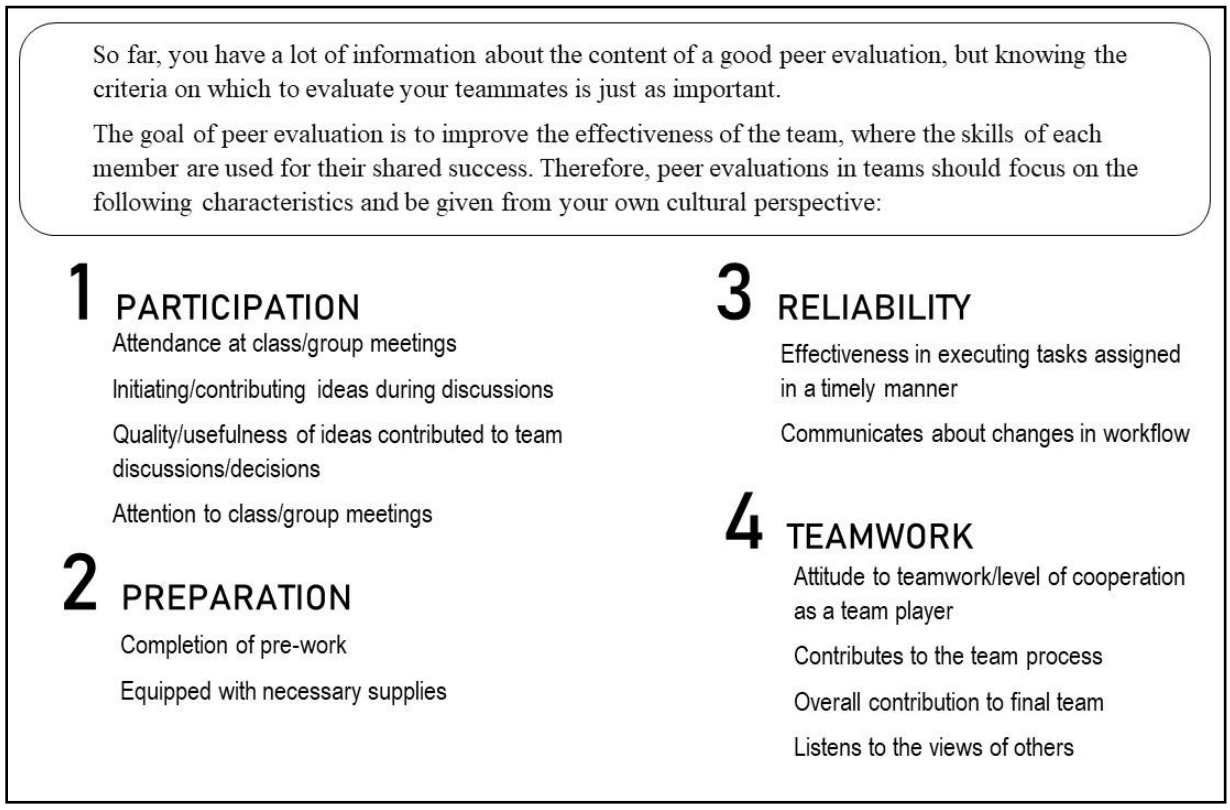

Figure 7. Items to be considered in peer assessments

Participants were then tasked with categorizing characteristics of teammates (e.g. “My teammate is considerably older than the rest of the class” or “My teammate tends to interrupt”) into “Should consider when evaluating” and “Should not consider when evaluating”. After completing the categorization, feedback was given on the best way to divide the characteristics.

Part One: Putting it All Together

In the final section of Part One, participants were presented with scenarios and asked to put what they had just learned together to select the most appropriate positive and constructive feedback for their teammate (Figure 8). These scenarios implemented scaffolding by initially presenting the participant with multiple choice options for feedback and then moving on to feedback written by the participant.

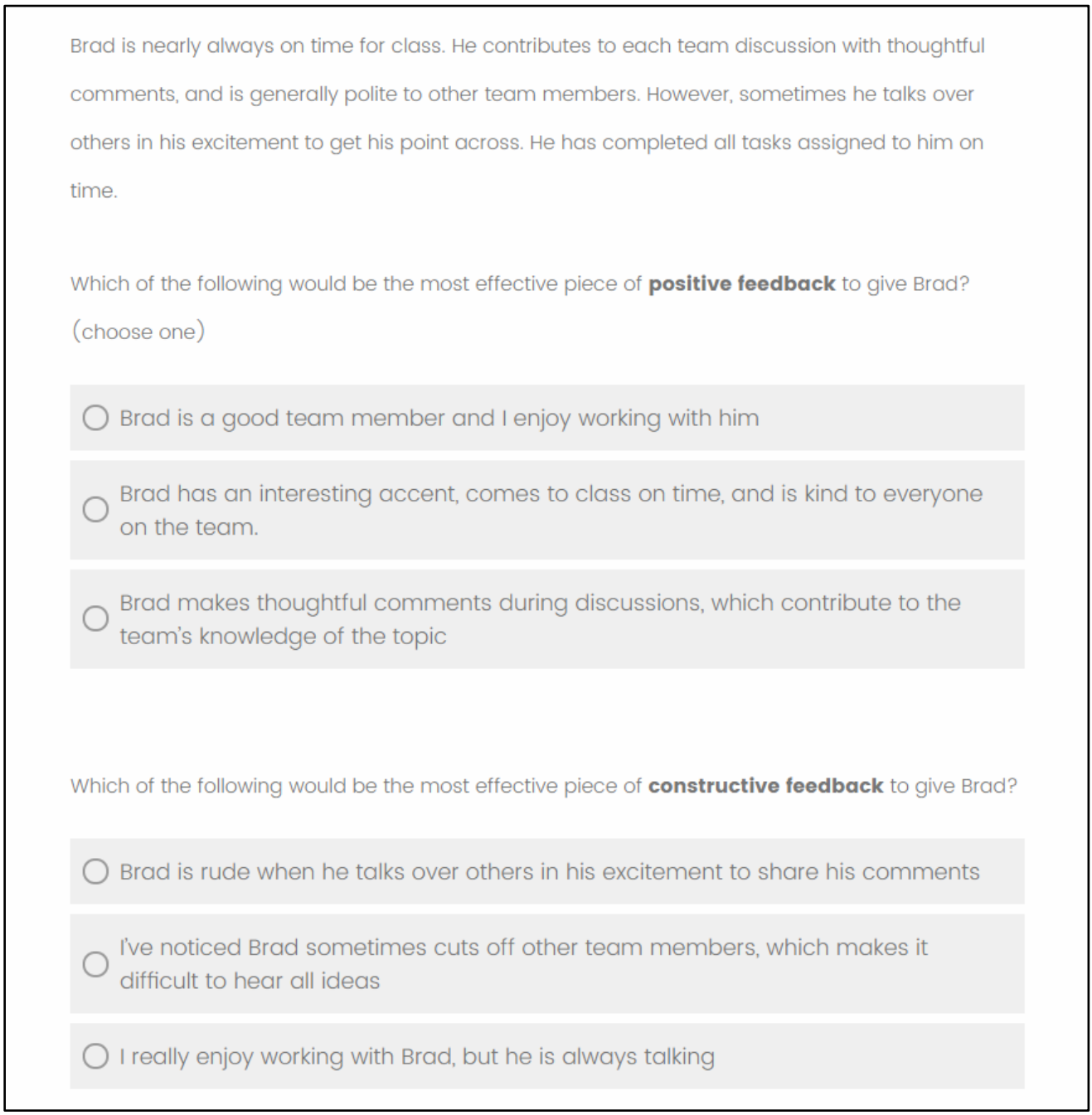

Figure 8. Scenario with positive and constructive feedback options

After each scenario, participants received feedback on their answers. If the participant chose a correct multiple choice answer, they were presented with rationale for why their choice

was correct. Similarly, if an incorrect answer was selected, the participant was shown the correct answer and its rationale, as well as why their chosen answer was incorrect (Figure 9).

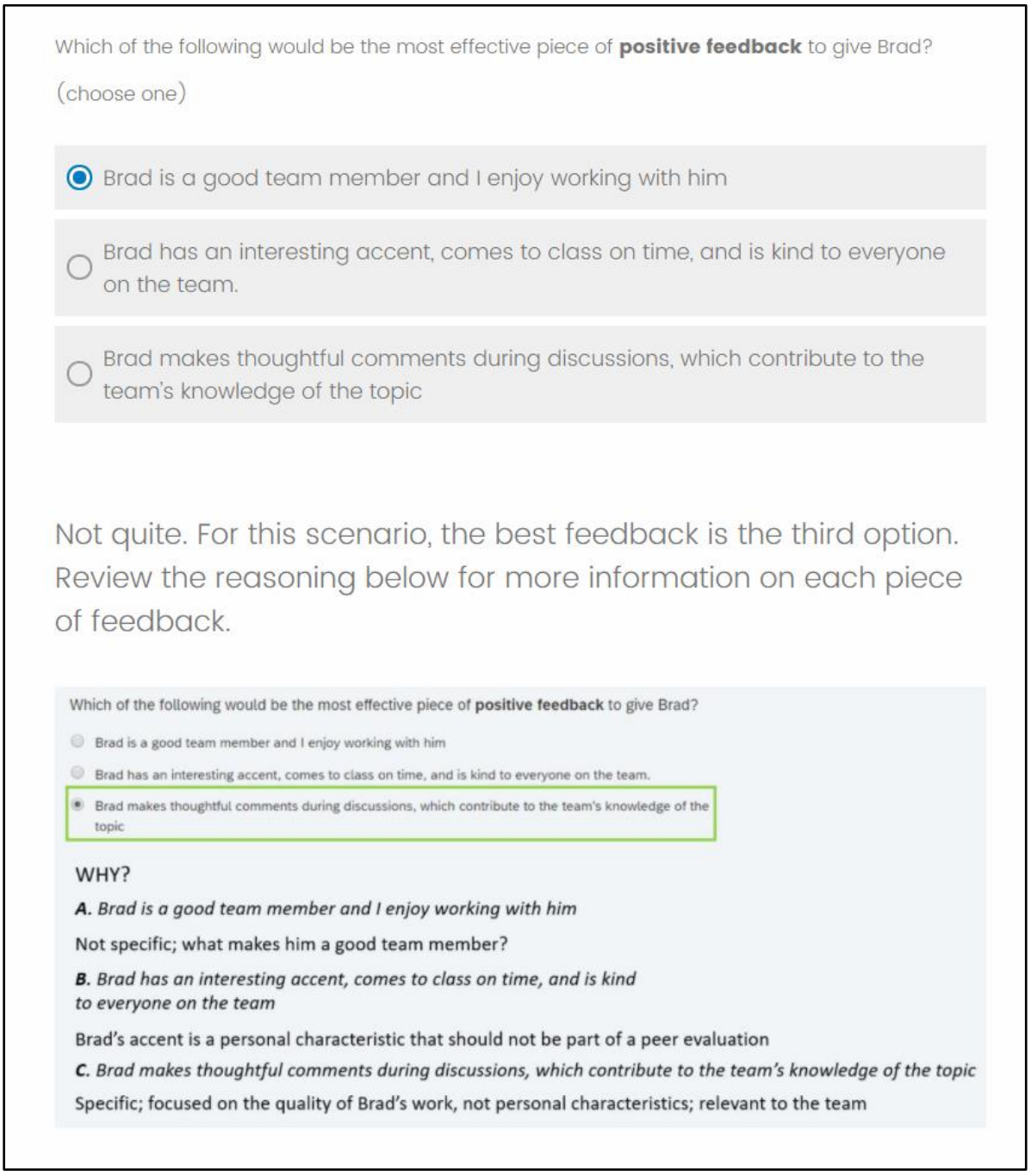

Figure 9. Feedback to participant on answer selection

Part Two: Assessing Peers Fairly

The design of the Part Two of the training was based upon a "prejudice habit breaking" model proposed by Patricia Devine and colleagues (2012). In this model there are three stages: awareness and consequences, techniques for mitigation, and time to practice (Devine et al., 1991; Devine et al., 2012). This model was selected for use in this training due to evidence that it can produce lasting reduction in implicit bias and its previous successful use in the higher education classroom environment (Devine, et al., 2012; Carnes, et al., 2015).

The first two sections of Part Two corresponded with the awareness and consequences stage in the Devine model. Students were reminded of one of the seven characteristics of helpful feedback: feedback should be usable and concerned with behavior over which the receiver has control, which was then connected to the idea of bias. Biased feedback often takes into account qualities the receiver cannot control, such as gender, race, or age (Wayland, et al., 2014). Concepts such as implicit bias were also defined.

Part Two: Microaggressions

Implicit bias was then connected to the concept of *microaggressions*. Previous work in this area (e.g. Boysen et al., 2009; Chapter 4) has shown that classroom and peer assessment biases often manifest as microaggressions (e.g. in-class comments about country of origin or peer assessment comments about an accent). Therefore recognition and mitigation of microaggressive behavior became the backbone of the training's content. Students were introduced to the concept of microaggressions and their consequences through two videos (Figure 10). The first video, *What Are Micro-Aggressions,* provides definitions and examples of what microaggressions look like (Catharsis Productions, 2017). The second video, *How are microaggressions like mosquito bites?*, uses moments of humor combined with reality to gently

introduce the viewer to the consequences of microaggressions while reducing discomfort (Fusion Comedy, 2016). Reducing discomfort is an important part of making training successful, as discomfort can lead to defensiveness, which in turn undermines the effectiveness of the training (Carter, Onyeador, and Lewis, 2020). *How are microaggressions like mosquito bites* also utilizes elements of the bias mitigation strategy perspective taking, which has been shown to be effective in reducing implicit bias (Todd, Bodenhausen, Richeson, and Calinsky, 2001). In this strategy, individuals take the perspective of another person experiencing bias and its consequences.

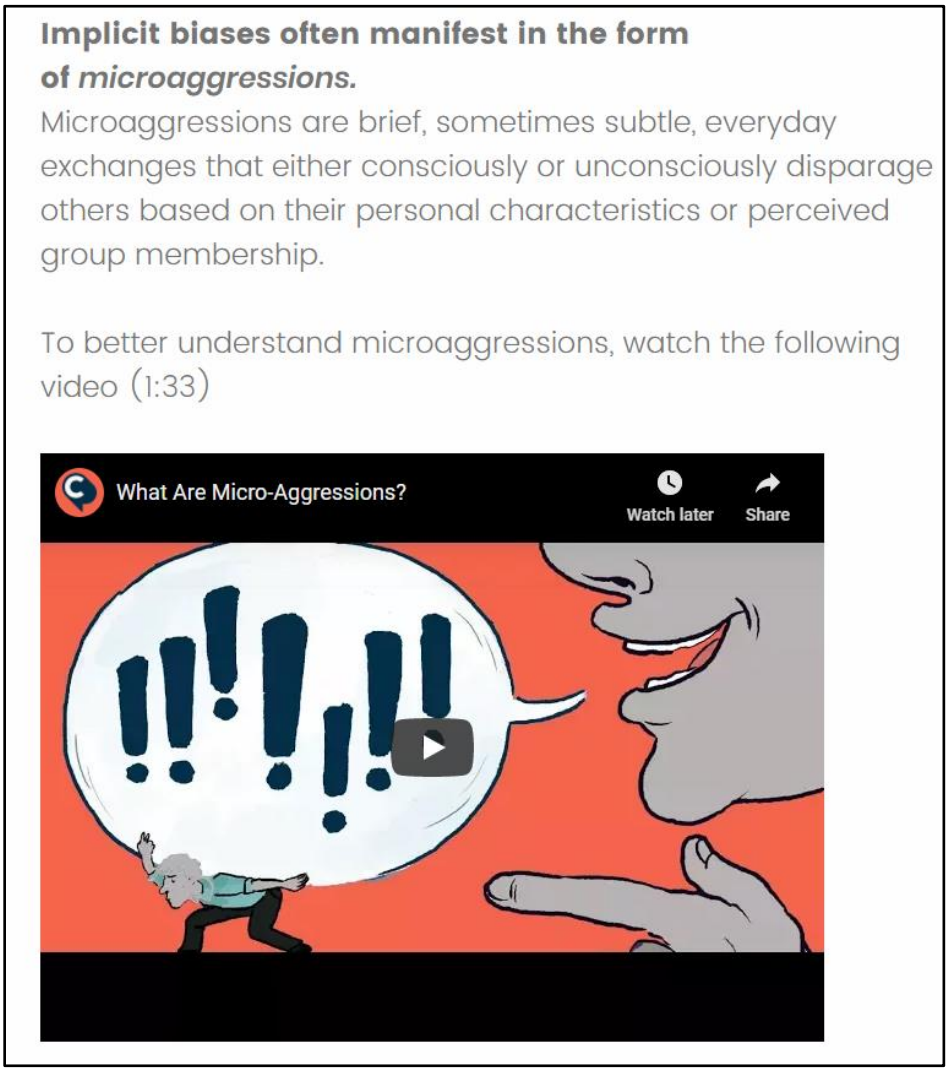

Figure 10. Introduction to microaggressions and video

Part Two Microaggressions and Peer Assessment

The last section of Part Two focused on the "techniques to mitigate bias" phase of the Devine model. In this phase, students were taught to identify microaggressions through matching activities, understand their hidden derogatory meanings, and rewrite microaggressive statements using unbiased language. This technique of identifying a biased response and rephrasing or replacing it is called *stereotype replacement.* Stereotype replacement is a bias mitigation technique which uses conscious rationalization to "retrain" oneself not to respond in a biased manner and has been shown effective at reducing implicit bias (Fine, Wendt, and Carnes, 2014; Monteith, 1993). Students then connected the concept of microaggressions back to peer assessment by categorizing potential peer feedback based on whether or not it contained a microaggressive statement (Figure 11).

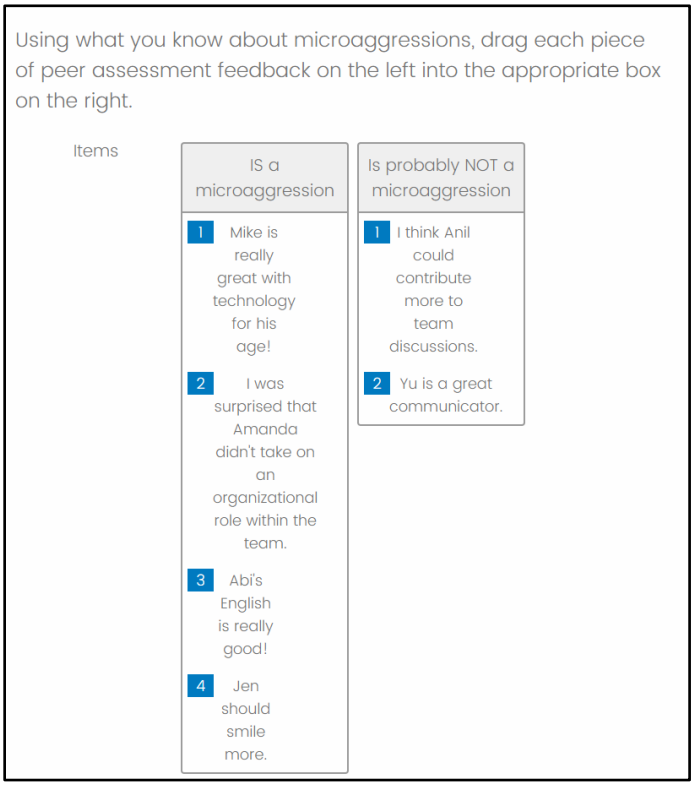

Figure 11. Categorization of potential peer assessment feedback

To reflect upon this categorization, students were also asked to take the role of a bystander and write about how they would respond to biased actions or statements in a team. Finally, participants were given a description of a team member and accompanying peer assessment comment and asked to identify microaggressive content (Figure 12).

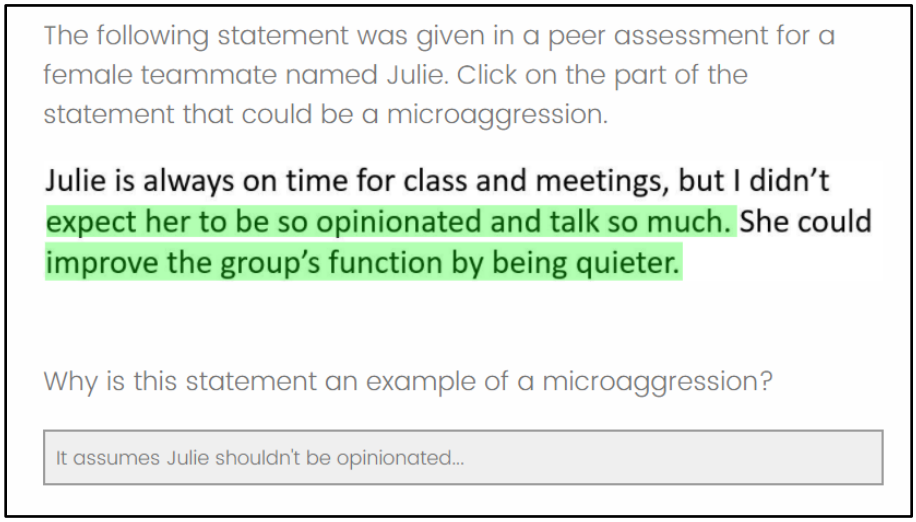

Figure 12. Peer assessment comment with possible microaggressive statements identified by highlight

After identifying the content, they then practiced rewriting the comment using all the information they had learned throughout the training. The third stage of the Devine training model, "time to practice" occurs as the semester progresses and students apply their knowledge.

Method

An Institutional Review Board (IRB) approved survey study (Appendix E) of the use and effect of peer assessment bias training was conducted in four Iowa State University classrooms.

Objective and Hypotheses

The objective of the study was to determine the effect peer assessment fairness training had on student perceptions of peer assessment fairness, attitudes toward peer assessment,

attitudes toward classroom teams, peer assessment score, and self-assessment score. The reception of the training was also studied.

Participants

Participation in the study was at the class level. Instructors were recruited for the study via email. Participating instructors were then asked to add a statement to their syllabus explaining that the class was taking part in a study of peer assessment, and that the students may be asked to complete surveys and trainings related to the peer assessment process. Four classes initially began the study, however due to the COVID-19 pandemic and the switch to online learning in April 2020, only one class, a 300-level Agronomy course, completed the study. Thirty students from this class completed the demographics portion of the study.

Fourteen participants identified themselves as women, 13 identified themselves as men, and three preferred not to state a gender. Twenty-five participants identified as White, one participant identified as Black, one participant identified as Hispanic, one participant identified as Pacific Islander, and two participants preferred not to answer. Twenty-nine participants were native English speakers, while one was not. Participants also identified the college with which they were affiliated (Table 19).

Table 19. Participant counts by academic college (N = 30)

College	**Participant Count**
Agriculture and Life Sciences	17
Human Sciences	6
Liberal Arts and Sciences	4
Engineering	3

Six sophomores, 13 juniors, and nine seniors took part in the study. Two students did not provide a class level. One participant reported being an international student, while 29 were not. Participant-reported GPAs averaged 3.3 (ranged from2.2-4.0).

Procedure

There were a total of nine activities in the project: a pre-survey, three self-assessments, three post-peer assessment surveys, the training, and a post-survey (Figure 13). All activities for the study were administered as assignments in Canvas which contained a link to the Qualtrics page for each item. Peer assessments were administered by the class instructor in Thinkspace. Participants were not divided into conditions (i.e. all students were assigned each activity as graded work), however individual participation in each activity varied.

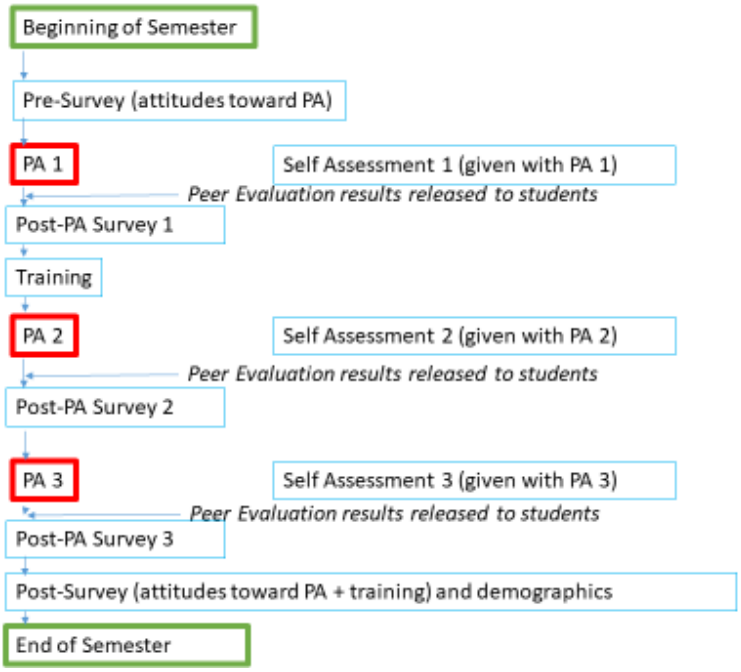

Figure 13. Timeline of activities in the study

Quasi-independent Variables

The study included two quasi-independent time variables: Study Time and PA Period. Study Time has two levels which correspond to the beginning of the semester (pre) and the end

of the semester (post) (green boxes in Figure 15). PA Period has three levels; one for each peer assessment period (red boxes in Figure 15).

Participants were not divided into conditions. However, some students chose not to complete the training. Therefore, the variable "training status" (yes/no) was used to analyze the data for differences between the students who did and did not complete training. Twenty participants completed the training while 10 did not. Selected demographics were also compared. The variable "gender" was simplified into two categories: Men and Women due to lack of participants in the "other" and "prefer not to say" categories. The remaining demographics (e.g. English speaker status) were not compared due to lack of participants.

Measures

The activities in the study may be broken down into measures by topic (Table 20).

Table 20. Measures used in the evaluation

Dependent Variable	**Metric**	**Units**	**Assessment Frequency**	**Activity**
Classroom team attitudes	4 survey questions	Likert scale 1-7	Study Time	Pre and post survey
PA attitudes and general fairness	8 survey questions	Likert scale 1-7, Likert scale 1-5 1 follow up free response	Study Time	Pre and Post survey
PA Fairness in specific class	5 survey questions	Likert scale 1-7, Likert scale 1-5 1 follow-up free response	Study Time (post-survey only)	Post survey
Reception of training	11 survey questions	Likert scale 1-5 3 follow up free response	Study Time (post-survey only)	Post survey
Peer Assessment	Michaelson method	0-15 points per team member 2 free response	PA Period	Peer Assessment
Self-assessment	Survey based on Michaelson method	0-15 points 2 free response	PA Period	Self Assessment
Fairness of PA Period	10 survey questions	Likert Scale 1-7, Likert scale 1-5 4 follow up free response	PA Period	Post Peer Assessment Survey
Demographics	10 survey questions	Multiple choice, free response	Study Time (post survey only)	Post survey

The pre-survey was used to establish baseline knowledge of students' attitudes toward working in teams, peer assessment, and any incidents of bias they might have perceived. The post survey was similar in structure to the pre-survey, but included additional sections on the training and demographics. Questions given in the pre and post surveys are divided into sections corresponding to the measures in Table 20.

Classroom Team Attitudes

Questions on classroom team attitudes were used to determine whether there was a change in student perceptions of classroom teams from the beginning to the end of the semester. The wording of some questions changed between the pre and post survey; both wordings are given in Table 21.

Table 21. Questions on classroom team attitudes

Pre-Survey Wording	Post-Survey Wording	Type
How much do you agree: I like working in teams on class projects	How much do you agree: I like working in teams on class projects	Likert (1-Strongly Disagree; 7-Strongly Agree)
How much do you agree: I avoid classes that involve teamwork?	How much do you agree: I avoid classes that involve teamwork?	Likert (1-Strongly Disagree; 7-Strongly Agree)
How much do you agree: When working on a team in a class or on a class project, I feel respected by my teammates	How much do you agree: When working on a team in a class or on a class project, I feel respected by my teammates	Likert (1-Strongly Disagree; 7-Strongly Agree)
How much do you agree: My classmates want to be on a team with me.	How much do you agree: My classmates want to be on a team with me.	Likert (1-Strongly Disagree; 7-Strongly Agree)

PA attitudes and General Fairness

Questions on attitudes toward peer assessment and general peer assessment fairness were used to determine if students' perceptions of peer assessment and its fairness changed from the

beginning to the end of the semester. Both pre and post survey question wordings are given in Table 22.

Table 22. Questions on attitudes toward peer assessment and general fairness

Pre-Survey Wording	Post-Survey Wording	Type
How much do you agree: I feel that the peer assessments I have received were fair	How much do you agree: Prior to this class, the peer assessments I received in other classes were fair	Likert (1-Strongly Disagree; 7-Strongly Agree)
How much do you agree: I feel that the peer assessments I have given were fair	How much do you agree: Prior to this class, the peer assessments I gave in other classes were fair	Likert (1-Strongly Disagree; 7-Strongly Agree)
How frequently have you felt unfairness in the peer assessments you have received?	Prior to this class, how frequently have you felt unfairness in the peer assessments you have received?	Likert (1-Never; 5-Always)
Follow-up: Please describe why you felt the peer assessment(s) you received were unfair.	*Follow-up:* Please describe why you felt the peer assessment(s) you received were unfair.	Free response
What characteristics or behaviors of your own could cause you to receive a lower peer assessment score?	In general, what characteristics or behaviors of your own could cause you to receive a lower peer assessment score?	Free response
What characteristics or behaviors of your teammates could cause you to give a lower peer assessment score?	In general, what characteristics or behaviors of your teammates could cause you to give a lower peer assessment score?	Free response
How much do you agree: I am confident I can assess my teammates fairly	How much do you agree: I am confident I can assess my teammates fairly	Likert (1-Strongly Disagree; 7-Strongly Agree)
How much do you agree: I am confident my teammates can assess me fairly	How much do you agree: I am confident my teammates in this class can assess me fairly	Likert (1-Strongly Disagree; 7-Strongly Agree)
How much do you agree: Peer assessment is beneficial to my learning	How much do you agree: Peer assessment is beneficial to my learning	Likert (1-Strongly Disagree; 7-Strongly Agree)

PA Fairness in Specific Class

Some questions on the fairness of peer assessments specifically in the class being studied were included only in the post survey. These questions may be found in Table 23.

Table 23. Items relating to PA fairness in the class enrolled in the study

Post-Survey Question	Type
How much do you agree: The peer assessments I received in this class were fair	Likert (1-Strongly Disagree; 7-Strongly Agree)
How much do you agree: The peer assessments I gave in this class were fair	Likert (1-Strongly Disagree; 7-Strongly Agree)
How frequently have you felt unfairness in the peer assessments you have received in this class?	Likert (1-Never; 5-Always)
Follow-up: Please describe why you felt the peer assessment(s) you received were unfair.	Free response
How much do you agree: I am confident my teammates in future classes can assess me fairly	Likert (1-Strongly Disagree; 7-Strongly Agree)

Reception of training

Questions on the reception of the training, its benefit, and methods of improvement were included only on the post survey. These questions are listed in Table 24.

Table 24. Items on reception of the training

Post-Survey Question	Type
How beneficial was the training?	Likert (1-Not beneficial at all; 5-Extremely beneficial)
What were the benefits of the training?	Free response
What was the most memorable part of the training?	Free response
Did you use the information you learned in the training when completing your peer assessments in this class?	Likert (1-Definitely did not use; 5-Definitely used)
Do you feel like your teammates used the information in the training when completing their peer assessments in this class?	Likert (1-Definitely did not use; 5-Definitely used)
Did you use the information you learned in the training when completing your peer assessments in other classes?	Likert (1-Definitely did not use; 5-Definitely used; N/A-None of my other classes used peer assessment)
Follow-up: What information did you use from the training?	Free response
What are your suggestions for improving the training?	Free response

Table 24 Continued	
Post-Survey Question	**Type**
Did your team discuss the training?	Likert (1-Definitely did not discuss or mention the training; 5-Discussed the training in depth)
Follow-up: What about the training did you discuss?	Free response
Follow-up: How beneficial was the discussion?	Likert (1-Not beneficial at all; 5-Extremely beneficial)

Peer Assessments

Peer Assessments were administered using the Michaelson method. In this method, students assign a numerical score to their teammates based upon the extent to which they believe their teammates contributed to the team as a whole (Michaelsen, 2002). A hallmark of this method, however, is that students may not assign everyone the same score. For example, if a student was dividing 40 points among four teammates, they would not be allowed to give each teammate 10 points. The peer assessments also included two free response questions on team members' contributions and how they could improve (Table 25).

Table 25. Items in the Michaelsen method self and peer assessments

Question	Type
How would you rate (your/this team member's) contribution to the team?	Numeric scale (0-15)
What is (your/this team member's) most important contribution to your team's effectiveness?	Free response
What is the most important thing (you/your team member) could change to improve your team's effectiveness?	Free response

Self-Assessments

The self-assessments, given with each peer assessment, were designed to be identical to the Michaelsen method peer assessments that the students completed. Traditionally, peer assessments in this method do not include ratings for oneself, however this component was

included in the study for the purpose of comparison with peer assessment scores. Self-assessment was included in this study in order to compare a student's view of their contribution with their team's view. The questions used in the self-assessment and peer assessment are located in Table 25.

Fairness of PA Period

The post-peer assessments, given after the students had the results of their peer assessment, were designed to gather the students' thoughts on the utility and fairness of the individual assessment. The questions used in the post-peer assessment are located in Table 26.

Table 26. Items relating to the fairness of the PA period

Question	Type
How accurate were the numerical scores you received on your peer assessment?	Likert (1-Extremely inaccurate; 7-Extremely accurate)
How accurate was the written feedback you received on your peer assessment?	Likert (1-Extremely inaccurate; 7-Extremely accurate)
Thinking only about this peer assessment, what was the most helpful comment you received from your teammates?	Free response
Why was this comment helpful?	Free response
Thinking only about this peer assessment, what was the least helpful comment you received from your teammates?	Free response
Why was this comment unhelpful?	Free response
How much do you agree: The peer assessments I received during this assessment period were fair	Likert (1-Strongly Disagree; 7-Strongly Agree)
Follow-up: Why?	Free response
How much do you agree: The peer assessments I gave during this assessment period were fair	Likert (1-Strongly Disagree; 7-Strongly Agree)
Follow-up: Why?	Free response
Was the written feedback you received from this peer assessment actionable?	Likert (1-Definitely not; 5-Definitely yes)
How likely are you to change a behavior based upon the feedback you received on this peer assessment?	Likert (1-Extremely unlikely; 5-Extremely likely)
Follow-up: What behavior will you change?	Free response

Table 26 Continued	
Question	**Type**
Follow-up: How will you change this behavior?	Free response

Demographics

Demographics were assessed only at the end of the study as part of the post survey. This was done to prevent stereotype threat (Steele & Aronson, 1995). Demographics questions are listed in Table 27.

Table 27. Demographic questions

Post-Survey Question	Type
What is your gender?	Multiple choice: Woman, man, other, prefer not to say
How would you describe yourself?	Multiple choice **(select all that apply):** White, Black or African American, Native American or Alaska Native, Asian, Native Hawaiian or Pacific Islander, Hispanic or Latinx, Other, Prefer not to answer
Is English the first language you learned to speak?	Yes/no
Follow-up: How many years have you been speaking English?	Free numeric response
Follow-up: How comfortable are you with speaking English?	Likert (1-Extremely Uncomfortable; 7-Extremely Comfortable)
Which college are you affiliated with?	Agriculture and Life Sciences, Business, Design, Engineering, Graduate, Human Sciences, Liberal Arts and Sciences, Veterinary Medicine
What is your major?	Free response
What class level are you?	Multiple choice: Freshman, Sophomore, Junior, Senior, Graduate
Are you an international student?	Yes/No
What is your GPA?	Select a number

Data Analysis

Free response survey data was analyzed for overall counts and themes. Survey items which utilized Likert scales were analyzed using a one-way or multiway ANOVA for comparison over time (pre/post survey or peer assessment periods one, two, and three), training status, and demographic variables. A significance level of $\alpha = .05$ was used throughout. Remaining demographic variables were not compared due to lack of students in each category.

Results

Classroom Team Attitudes

The comparisons of team-based class attitudes by Study Time (pre and post) are summarized in Table 28. There were no significant differences over time for any items.

Table 28. Comparison of classroom team attitudes. * indicates a significant result

Question		**Pre**		**Post**			
	N	*M*	*SD*	*M*	*SD*	*F*(1,N-1)	*p*
I like working in teams on class projects	29	3.7	1.1	4.0	.9	0.15	.69
I avoid classes that involve teamwork	29	3.6	1.5	4.2	1.4	1.35	.25
When working on a team in a class or on a class project, I feel respected by my teammates	29	5.6	1.0	5.2	1.1	1.15	.29
My classmates want to be on a team with me.	29	5.2	.9	5.1	1.2	0.001	.98

The means of each classroom team attitude item are compared by Study Time visually in Figure 14.

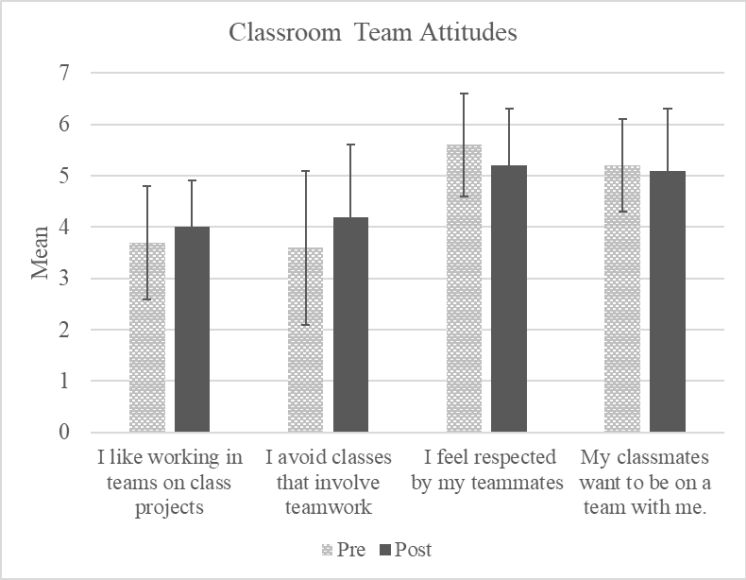

Figure 14. Classroom team attitudes by Study Time (pre/post). Bars indicate standard deviation

The results for the effect of gender on classroom team attitudes are summarized in Table 29. There were no significant interactions between gender and Study Time (pre/post).

Table 29. Effect of gender on classroom team attitudes. * indicates a significant result

Question		Men		Women			
	N	*M*	*SD*	*M*	*SD*	*F*(1, N-1)	*p*
I like working in teams on class projects	57	4.0	1.2	4.0	1.0	0.04	.89
I avoid classes that involve teamwork	57	3.9	1.0	4.9	1.0	0.008	.93
When working on a team in a class or on a class project, I feel respected by my teammates	57	5.7	1.3	5.2	1.1	3.08	.042* (*d* = .44)
My classmates want to be on a team with me.	57	5.7	1.4	4.7	1.2	8.66	.005* (*d* = .71)

The means of each classroom team attitude item are compared by gender visually in Figure 15. Significant differences are noted with *.

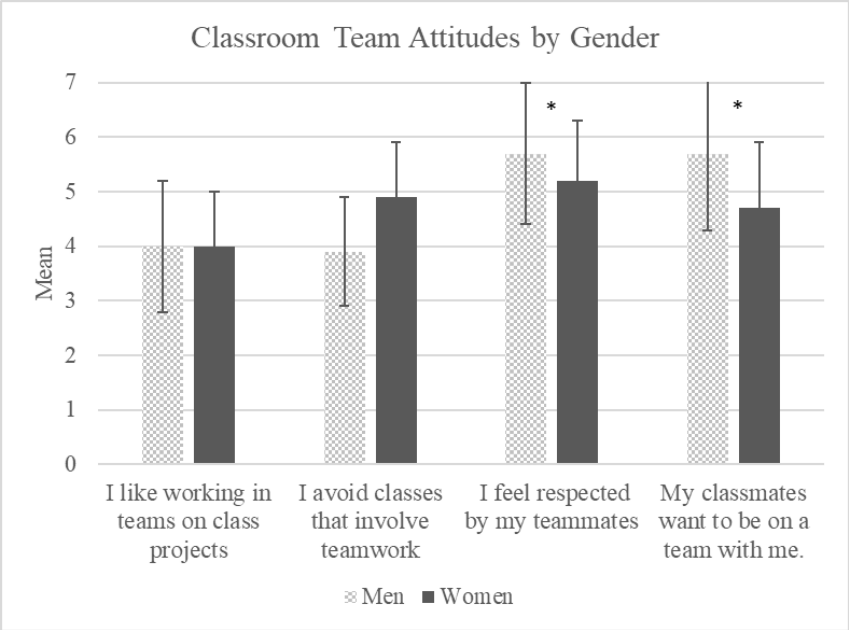

Figure 15. Classroom team attitudes by gender. Bars represent standard deviation; * indicates means which are significantly different

PA Attitudes and General Fairness

The comparison of peer assessment attitudes and beliefs by Study Time (pre and post) are summarized in Table 30. Two items, *the peer assessments I received in other classes were fair* and *the peer assessments I gave in other classes were fair* achieved significance.

Table 30. Peer assessment attitudes and general fairness by Study Time (pre/post)

Question		**Pre**		**Post**			
	N	*M*	*SD*	*M*	*SD*	*F*(1,N-1)	*p*
Prior to this class, the peer assessments I received in other classes were fair	27	5.9	1.3	5.1	1.2	6.77	.012* (*d* = .61)
Prior to this class, the peer assessments I gave in other classes were fair	28	6.1	1.2	5.5	1.2	4.89	.032* (*d* = .5)
Prior to this class, how frequently have you felt unfairness in the peer assessments you have received?	28	2.1	1.12	1.9	1.3	0.21	.65
I am confident I can assess my teammates fairly	26	5.9	1.3	6.2	1.4	1.77	.19

Table 30 Continued							
Question		**Pre**		**Post**			
I am confident my teammates in this class can assess me fairly	28	5.7	1.33	5.3	1.2	2.82	.09
Peer assessment is beneficial to my learning	28	4.4	1.4	4.2	1.3	0.21	.65

An analysis of peer assessment attitudes and fairness by training status was also conducted. There was a significant difference in confidence in assessing teammates fairly where the training group (M = 6.2, SD = 1.3) indicated higher confidence than the non-training group (M = 5.6, SD = 1.2); $F(1, 54) = 2.16$, p = .038, d = .53. There was a significant difference in confidence teammates could rate the participant fairly where the training group (M = 5.9, SD = 1.3) indicated higher confidence than the non-training group (M = 5.4, SD = 1.0); $F(1, 54)$ = 1.60, p = .035, d = .47 . There was a significant difference in confidence teammates in future classes could rate the participant fairly where the training group (M = 5.7, SD = 1.4) indicated higher confidence than the non-training group (M = 5.0, SD = 1.1); $F(1, 54) = 3.69$, p = .009, d = .62 There were no significant interactions between training and Study Time (pre/post). Comparisons of fairness by training status are shown visually in Figure 16.

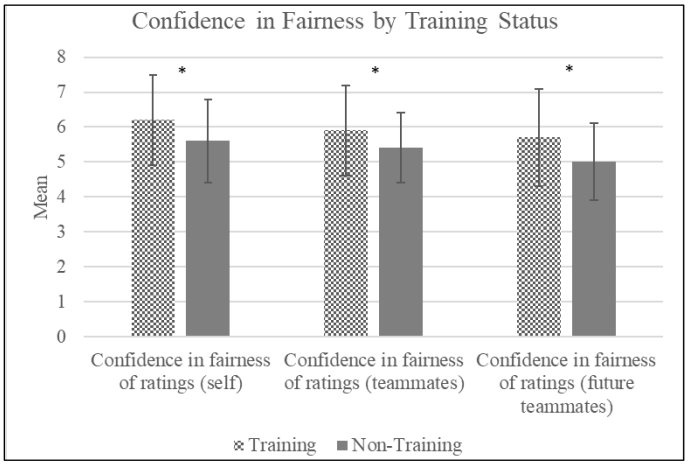

Figure 16. Confidence in fairness by training status. Bars represent standard deviation; * indicates means which are significantly different

The results for the effect of gender on peer assessment attitudes and fairness are summarized in Table 31. There were no significant interactions between gender and Study Time (pre/post).

Table 31. Effect of gender on peer assessment attitudes and fairness. * indicates a significant result

Question		**Men**		**Women**			
	N	*M*	*SD*	*M*	*SD*	*F*(1, N-1)	*p*
Prior to this class, the peer assessments I received in other classes were fair	27	5.7	1.1	5.4	1.4	0.83	.37
Prior to this class, the peer assessments I gave in other classes were fair	28	6.2	1.1	5.4	1.4	7.38	.009* (*d* = .64)
Prior to this class, how frequently have you felt unfairness in the peer assessments you have received?	28	1.8	1.6	2.2	1.4	2.24	.14
I am confident I can assess my teammates fairly	26	6.3	1.9	5.8	1.7	3.45	.069
I am confident my teammates in this class can assess me fairly	28	6.0	1.6	5.6	1.8	1.82	.19
Peer assessment is beneficial to my learning	26	4.6	1.8	3.9	2.2	1.73	.20

PA Fairness in Specific Class

There were no significant differences in perceived fairness of the assessments received in the participant's class between the training group (M = 5.6, SD = 1.2) and the non-training group (M = 5.6, SD = 1.5); $F(1, 28) = .022$, $p = .88$. There were no significant differences in perceived fairness of the assessments given by the participant in the participant's class between the training group (M = 6.0, SD = 1.0) and the non-training group (M = 5.8, SD = 1.2); $F(1, 28) = .35$, p =

.56. There were no significant differences in perceived frequency of bias in the participant's class between the training group (M = 2.0, SD = 1.0) and the non-training group (M = 2.0, SD = 1.0); $F(1, 29) = 0.00$, $p = 1.0$.

There was a significant difference in perceived fairness of the assessments received in the participant's class where men (M = 5.9, SD = 1.0) indicated greater fairness than women (M = 5.4, SD = 1.2); $F(1, 26) = 4.45$, $p = .045$, $d = .50$. There were no significant differences in perceived fairness of the assessments given by the participant in the participant's class between men (M = 6.3, SD = 2.1) and women (M = 5.7, SD = 1.8); $F(1, 26) = 3.02$, $p = .094$. Comparison of perceived fairness by gender is shown visually in Figure 17.

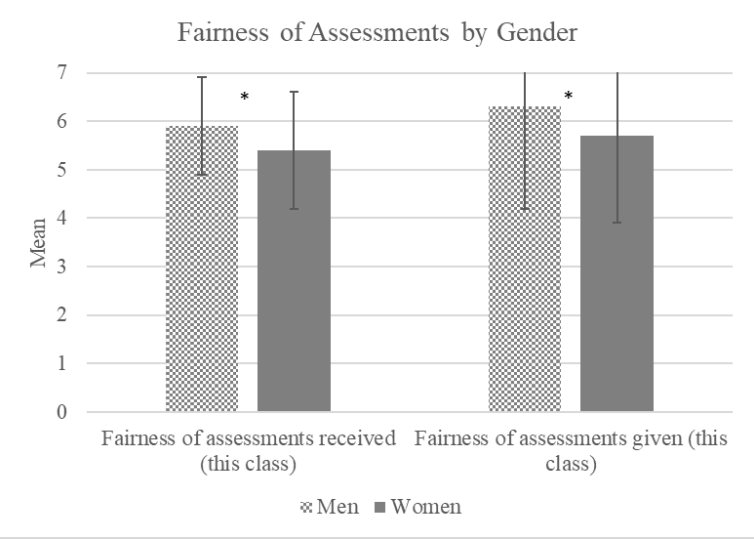

Figure 17. Fairness of peer assessments received by gender. * indicates means which are significantly different

There was a significant difference in perceived frequency of bias in the participant's class where men (M = 1.4, SD = 1.1) indicated bias occurring less frequently then women (M = 2.5, SD = 1.2); $F(1, 26) = 12.65$, $p = .0015$, $d = .98$.

Reception of Training

Twenty of the 30 participants participated in the training. The mean score for the benefit of the training was 2.71 (out of 5, *SD* = 1.12). Specific benefits of the training listed were: knowing what criteria to evaluate teammates against, examples of constructive comments, recognition of the importance of feedback from peers, learning what could be hurtful to a teammate and why, and instructions on how to craft a constructive comment. The mean score for using the information in the training when completing peer assessments was 3.62 (out of 5, *SD* = 1.17) and the mean score for feeling teammates used the information in the training was 3.2 (out of 5, *SD* = .98). Participants who were enrolled in other classes using peer assessment (N = 12) were asked to evaluate whether they used the training material in other peer assessment classrooms (*M* = 1.76; *SD* = 1.80). Students' comments indicated that the information they used from the training included: how to give thoughtfully worded comments, how to make constructive comments actionable, and avoiding personal bias. The mean score for having discussed the training as a team was 1.57 (out of 5, *SD* = .80). Students who indicated their team had discussed the training mentioned discussing how it could be helpful and its length. Suggestions for improving the training included: reducing length, making reference material more easily accessible, moving it to an in-person class, more interaction, and removing it altogether. Comparisons of items relating to training reception is shown visually in Figure 18.

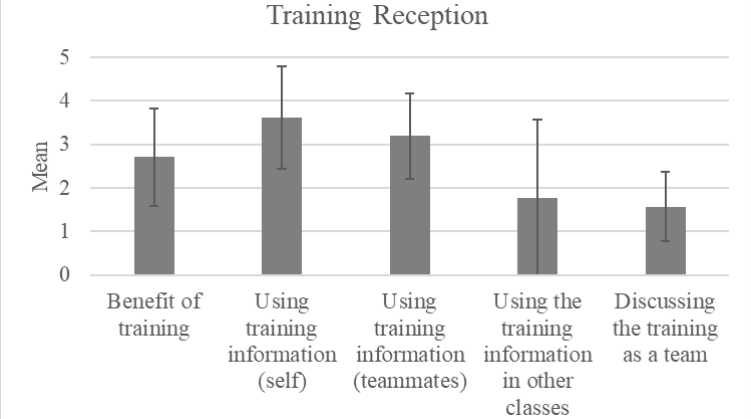

Figure 18. Training reception. Bars indicate standard deviation

Peer Assessments and Self Assessments

Peer assessment and self-assessment scores were significantly different where self-assessment scores (M = 10.9, SD = 1.6) were significantly higher than peer assessment scores (M = 10.0, SD = 1.5); $F(1, 250) = 15.97$, $p<.001$, $d = .58$. This comparison is presented visually in Figure 19.

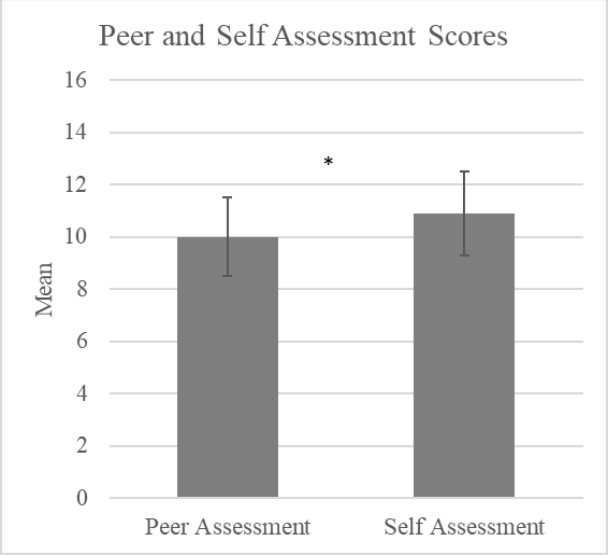

Figure 19. Peer and self assessment scores. Bars represent standard deviation; * indicates means which are significantly different

There were no significant differences in peer assessment scores between any of the assessment periods. There were no significant differences in self-assessment scores for any of the assessment periods. There was a significant difference for the peer assessment score received where women (M = 10.4, SD = 1.2) received significantly higher peer assessment scores than men (M = 9.7, SD = 1.1); $F(1, 134) = 9.61$, $p = .002$, $d = .63$. There were no significant differences in self-assessment score by gender. There were no significant interactions between gender and PA Period on peer assessment score or self-assessment score.

Fairness of PA Period

Fairness of specific peer assessment was analyzed for the three levels of PA Period: period one, period two, and period three. There was a significant difference in perceived accuracy of the numeric feedback received where accuracy was lower in assessment period two (M = 5.3, SD = .9) than in assessment periods one (M = 5.9, SD = 1.3, $p = .03$, $d = .52$) and three (M = 5.8, SD = 1.2, $p = .031$, $d = .47$); $F(2, 108) = 3.58$, $p = .031$. There were no significant differences in perceived accuracy of the written feedback for any of the assessment periods (P1: M = 5.8, SD = 1.3; P2: M = 5.6, SD = 1.1; P3: M = 5.8, SD = .9); $F(2, 108) = .39$, $p = .67$. There were no significant differences in perceived fairness of the peer assessments received for any of the assessment periods (P1: M = 5.9, SD = 1.9; P2: M = 5.8, SD = 1.4; P3: M = 5.7, SD = 1.5); $F(2, 108) = .45$, $p = .64$. There was a significant difference in perceived fairness of the peer assessment given where fairness was lower in assessment period two (M = 4.9, SD = 1.0) than in assessment periods one (M = 6.0, SD = 1.1, $p < .001$, $d = 1.06$) and three (M = 5.9, SD = 1.2, $p < .001$, $d = .91$); $F(2, 107) = 10.72$, $p < .001$. There were no significant differences in feedback being actionable for any of the assessment periods (P1: M = 3.84, SD = 1.3; P2: M = 3.6, SD = 1.4; P3: M =3.4, SD = 1.3); $F(2, 107) = .44$, $p = .65$. There were no significant differences in

plans to change a team behavior for any of the assessment periods (P1: *M* = 3.6, *SD* = 1.1; P2: *M* = 3.5, *SD* = 1.0; P3: *M* =3.4, *SD* = 1.7); *F*(2, 75) = .155, *p* = .86. These results are summarized visually in Figure 20.

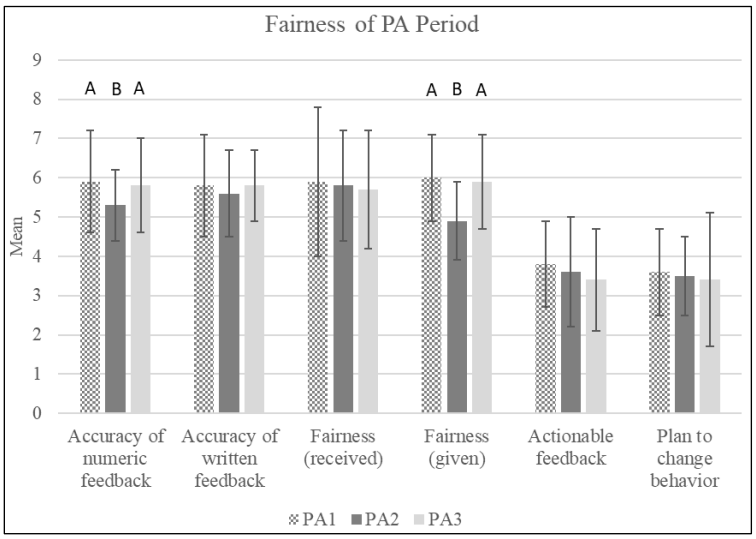

Figure 20. Fairness of PA period. Bars represent standard deviation; Items not connected by the same letter are significantly different

There were no significant differences in perceived accuracy of the numeric feedback received between men (*M* = 5.7, *SD* = 1.2) and women (*M* = 5.6, *SD* = 1.6); *F*(1, 107) = .16, *p* = .69. There were no significant differences in perceived accuracy of the written feedback for between men (*M* = 5.6, *SD* = 1.9) and women (*M* = 5.9, *SD* = 1.3); *F*(1, 107) = 1.56, *p* = .214. There were no significant differences in perceived fairness of the peer assessments received between men (*M* = 5.9, *SD* = 1.2) and women (*M* = 5.8, *SD* = 1.6); *F*(1, 107) = .64, *p* = .43. There were no significant differences in perceived fairness of the peer assessments given between men (*M* = 5.7, *SD* = 1.3) and women (*M* = 5.5, *SD* = 1.2); *F*(1, 107) = .44, *p* = .51. There was a significant difference in feedback received being actionable, where women (*M* = 3.9, *SD* = 1.0) found their feedback to be more actionable than men (*M* = 3.5 *SD* = 1.2); *F*(1, 107) = 5.02, *p* = .027, *d* = .41. There were no significant differences in plans to change a team

behavior based on feedback received for men (M = 3.4, SD = 1.0) and women (M = 3.6, SD = 1.8); $F(1, 75) = .304$, $p = .56$. There were no significant interactions between gender and PA Period for any of the items.

Discussion

Student perceptions of peer assessment fairness did not increase with each peer assessment period. While there was no change in perception of fairness in the participant's current class, there was an effect for recognition of unfairness (both in assessments given and received) in previous classes. This could be an indication that the training was successful in terms of the first phase of the Devine et al. (2012), model of bias reduction: Attention to bias and recognition of consequences. In this phase, an individual is made aware of the issue of bias and learns to recognize it when it occurs. Demonstrating recognition of past unfairness, both in the peer assessments given and received, represents an important milestone in understanding how to effectively and fairly assess peers as it shows that students are becoming better able to recognize bias and have completed the first step in the Devine et al (2012) process. In the Devine model, the next step would be to learn strategies to mitigate biased actions and put them to work. These strategies are taught in the current training, but could be approached differently in future iterations. Evidence of learning and using these strategies could come in the form of increased perceptions of fairness over time as well as increased confidence in students' rating skills.

In this study, 10 students did not complete the training. When analyzing confidence data with regard to the quasi-independent training variable, the results indicated that students in the training group reported higher confidence in their rating ability and the rating ability of their peers. Therefore, while confidence did not increase with time, it was higher for those who received training. This could indicate that the training was effective in increasing confidence for

those who took it. However, this effect could also be due to self-selection bias (Heckman, 1990) or the personal motivations of those who chose to complete the training (i.e. students who are unwilling to skip a class assignment may have higher confidence in their rating skills regardless of training).

There were no significant differences in attitude toward classroom teams between the beginning and the end of the study. In many classes that use teams, peer assessment is essential (Cestone, Levine, and Lane, 2008). Student attitudes toward teams in team based learning classes have been predicted by level teamwork efficiency, cooperation, and peer assessment (Ekimova & Kokurin, 2015). It was hypothesized that by training students on the peer assessment process, they would also learn to be more effective teammates, thereby increasing positive attitudes toward classroom teams as a whole. This outcome was not observed in the current study. This could suggest that attitudes toward peer assessment and attitudes toward classroom teams are less connected or that other predictors of attitude toward teams (e.g. teamwork efficiency) affected the results more than peer assessment and training. This could also suggest that the training in its current form does not have a strong enough effect to change student attitudes outside of peer assessment. Attitudes toward classroom teams have been improved through the use of team building exercises, however, so the incorporation of these exercises and training could be considered in future work (Ekimova & Kokurin, 2015).

In the results for peer assessment attitudes and general fairness, men indicated a higher level of perceived fairness for their ratings given in previous classes than women. However, in terms of ratings by gender, the results of the large-scale data analysis presented in Chapter 4 show that men give lower scores overall and rate women significantly lower than other men – an indication that their ratings may not be as fair as they perceive. Once again referencing Chapter

4, the data analysis showed that women receive significantly lower peer assessment scores than men. However, in this class, the opposite was true where women received higher peer assessment scores for each assessment period. In the literature, this result is often explained by GPA (women tend to have higher GPAs) or the gender connotation of the class (whether the class is considered a "masculine" or "feminine" topic) (May and Gueldenzoph, 2006; Baker, 2008). Both conclusions about fairness and peer assessment score by gender could have been influenced by the much lower number of participants in the present study than in any of the studies described in Chapter 4. Further, the discrepancy between the effect of gender on peer assessment scores observed in Chapter 4 and this study could be explained by investigating the gender association of the class. Results for the reception of the training show moderate scores for benefit and use, but a low score for discussing the training among teammates. While encouragement to discuss the training among teammates was not included in the training, such a discussion of course activities is expected in a team-based learning classroom and could increase retention of the material (McInerney & Fink, 2003). To increase potential for team discussion and retention of material, the training could be included in a Team Based Learning application exercise. Actionable student suggestions for improving the training were to make it shorter, increase interactivity, and provide easy access to reference material. These suggestions will be taken into consideration in the future version of the training.

The use of training and non-training conditions was unintentional. However, as the training was an assignment in a classroom, students were free to ignore it if they were willing to receive a lower grade. Students who chose not to complete the training did not provide reasons, therefore it is difficult to know their motivations. Unlike intentionally dividing participants into conditions, the researchers had no control over groupings and could not ensure they were

representative. This lack of control brings into question the validity of the significant results for training condition. It is possible that the attitudes that motivated some students to ignore the training influenced their survey responses more than the lack of training itself.

Additionally, students underwent a dramatic shift at midterm due to the Covid-19 outbreak in 2020. The class shifted from completely in-person to completely online in a matter of days over spring break. While the teamwork aspect of the course remained intact, shifting team interactions online can present challenges. Individual responses to these challenges may vary and could have affected the results of the surveys and peer assessments in the second half of the semester.

The results of this work, specifically the lack of significant change in perceptions of fairness and confidence, highlight the need for training refinement and deployment in a larger number of classes. This formative evaluation wasused as a starting point for gathering further requirements to increase the effectiveness of the training. These requirements were gathered through focus groups with stakeholders: students, instructors, and online learning professionals, and will be used to redesign the training for summative evaluation. While the quasi-independent training variable provided opportunities for analysis, in the summative evaluation, it will be essential that all students complete the training.

Conclusion

In this chapter, an initial iteration of training to address this issue was described and evaluated. While both bias and assessment have been targets of previous training endeavors, this study combines lessons learned from both into a new training methodology. The results of this evaluation of the training do not show definitive evidence of improved perceptions of peer assessment fairness, student confidence in their rating abilities, or attitudes toward classroom

teams. However, they do indicate steps in that direction. For example, data from the pre- and post- surveys indicate an increased awareness of biased peer assessments, which corresponds to the first step in the Devine model of bias mitigation (Devine, et al., 2012). With refinement, this training could improve peer assessment outcomes for students while bolstering instructors' confidence in their use of peer assessment. In Chapter 6, the process of refinement will continue through the use of focus groups to gather further requirements for the training. These requirements will then inform the redesign of the final training in Chapter 7.

CHAPTER 6. FOCUS GROUPS ON BIAS AND TRAINING

Research Objectives

Three focus groups were conducted with online learning professionals, instructors, and students, respectively. The objective of the online learning focus group was to understand best practices for online content delivery and receive feedback on the first iteration of the training. The instructor focus group targeted a deep dive into the instructor perspective on bias in the classroom and peer assessments, as well as feedback on the first iteration of the training. Similarly, the objective of the student focus group was to better understand how students perceive bias in the classroom and peer assessments, as well as gather feedback on the training. Data gathered from these focus groups was utilized to develop requirements for the revised training to ensure it meets online content delivery guidelines and addresses the needs and concerns of students and instructors. The data gathered from these focus groups informs RQ1 by fostering deeper discussion of bias in the classroom and peer assessment than can be gathered through the use of surveys alone. Focus group data also informed RQ3 through feedback on the first iteration of the training and requirements for further iterations.

Method: Online Learning, Student, and Instructor Focus Groups

Three Institutional Review Board (IRB) approved (Appendix C, Appendix D) Focus Groups of classroom and peer assessment bias as well as recommendations for training were conducted.

Participants

Participants for the study were recruited from Iowa State University through email. Potential participants for the instructor focus group were selected based on their response to a question on their survey (Chapter 4) asking if they would be willing to participate in future

research on the topic. All three participants in the instructor focus group were women. One participant was affiliated with the college of Human Sciences, one participant was affiliated with the college of Liberal Arts and Sciences, and one participant was affiliated with the college of Engineering. Each participant in the instructor group was an instructor at Iowa State University who had used peer assessment for at least five years.

Potential participants for the online learning focus group were selected based upon involvement with online learning at Iowa State University. All three participants in the online learning focus group were women. Each participant in the online learning focus group was a professional working with online course delivery design at Iowa State University.

Potential participants for the student focus group were selected based on their response to a question on their survey (Chapter 4) asking if they would be willing to participate in future research on the topic. All three participants in the student group were undergraduate women age 18 or older. Each participant in the student focus group had taken at least one class where peer assessment was used.

Procedure

The procedure for each focus group was the same, and therefore this section will describe the methods for all three groups simultaneously. Each participant received a link to an electronic consent form. After completing the consent form, the participant joined the focus group by clicking a link that led them to a WebEx meeting. Before beginning, participants were reminded that the focus group was not being recorded, their identity would not be shared in results, and that they should refrain from sharing the identity of fellow participants outside of the meeting.. Before answering questions directly pertaining to the training, participants were shown an overview of the training as well as screenshots and videos via screen sharing in the meeting.

Follow-up questions were asked where needed or as dictated by the direction of the discussion. Questions were delivered to the whole group at once verbally and by sharing a screen and were open for discussion. While participants were speaking, a researcher took notes on their responses.

Focus Group Questions

The questions posed in each focus group were categorized by topic. The first group of questions focused on biases that may have been observed or perceived in the classroom and in peer assessments (Table 32).

Table 32. Focus group questions on bias in the classroom and peer assessment

Question	Online learning	Instructors	Students
Have you experienced acts of bias in your classroom/in the classroom?		x	x
Have you noticed bias in peer assessments you have given/received?		x	x
In what ways do you think peer assessments could be biased?			x
Are you concerned about the fairness of peer assessments?			x

The second group of questions were focused on the idea of peer assessment training, but posed before the training was shown (Table 33). This was done to avoid priming participants' responses (Pashler, Coburn, & Harris, 2012).

Table 33. Focus group questions on training implementation

Question	Online learning	Instructors	Students
Do you use any kind of training for the students before their peer assessments?		x	
What are the barriers instructors might face when implementing a new training in their classrooms?	x	x	
How can the process of implementing the training be made easier for instructors?	x	x	
What are best practices for implementing a training in an online learning environment?	x		

Table 33 Continued			
Question	**Online learning**	**Instructors**	**Students**
Are there resources we should know about or review?	x		
What else do we need to know?	x		

The third group of questions were focused on the initial training described in Chapter 5 (Table 34). Questions in this group were crafted to garner feedback to be incorporated into the final version of the training. These questions were posed after participants had viewed the training materials.

Table 34. Focus group questions on initial peer assessment fairness training

Question	**Online learning**	**Instructors**	**Students**
After looking at the training, what comes to mind immediately?	x	x	x
Do you think this training would be helpful? Do you think your students would find the training useful?		x	
Do you think this training would be helpful or useful?			x
Given the feedback we have received from students in the first iteration of the training and your knowledge of instructional design, what are your suggestions?	x	x	
If any of the instructors have already participated in the study: What was your experience with the study and the training? What would you suggest be done differently?		x	
How would knowing your classmates had to complete this training impact how you thought about peer assessments?			x
What are your suggestions for improving the training?	x	x	x
Would creating Canvas modules for each part of the study and the training be helpful?	x	x	

Data Analysis

Notes on participant responses were analyzed for themes, which were then represented using an affinity diagram. An affinity diagram is used to organize large amounts of data into natural relationships (Widjaja & Takahashi, 2016). This process encourages collaboration while helping to diagnose issues and discover commonalities (Rowlands & Price, 2009). The affinity diagram was created by using the online tool, LucidSpark. To create the diagram, three researchers viewed electronic "cards", each containing a theme. All three researchers were US-born and White. Two researchers were women while one was a man. Two researchers were tenured faculty with experience giving peer assessments to their students and research interest in peer assessment and fairness. One researcher was a PhD candidate with experience in college teaching and research interest in peer assessment and fairness. The researchers silently arranged the cards into groupings. Researchers were able to rearrange cards placed by another researcher. Once all cards were arranged, the researchers discussed and gave each grouping a label, rearranging cards as necessary. The resulting affinity diagram was then used to better understand bias and inform requirements for further iterations of the training.

Results

Notes from the focus groups were decomposed into 109 theme cards which were organized into 20 categories. The resulting affinity diagram is shown in Figure 21. Cards in green are grouping labels while purple cards are student themes, red cards are instructor themes, and yellow cards are online learning professional themes.

Figure 21. Affinity diagram. Cards in green are grouping labels

For utility, grouping labels themselves can be categorized (Table 35). These categories facilitate interpretation of cards across groups.

Table 35. Categorization of grouping labels

Training Needs	Peer Assessment Process	Bias
• Why I want to use the training in my class • Instructor functional concerns before starting the training • Is the training relevant and useful to my class? • Instructor support for training • How to get students to take the training • Support for training beyond the classroom (development of professional skills) • Student needs in training (functional requirements)	• How to make the peer comments more fair • Assessment process description • Rules for peer assessment implementation • Team composition strategies • Barriers to giving constructive feedback • Are students being prepared/trained for peer assessment? • Anonymity	• Friendship biases, student relationships impacting peer assessment scores • Looking for biases • Overt biases that have been noticed • Things are fine in peer assessment • Assessment based upon perceived contributions • Fears and anxieties about bias conversations

From these categories and their associated cards, requirements for peer assessment training were extracted. To do so, the cards associated with the "Training Needs" column were examined. When examining the cards, focus was placed on actionable needs. For example, the card "do you need another piece of tech for the training?" yielded the requirement that training not require additional technology or platforms. Likewise, the cards "I would hate if the training was one giant paragraph" and "If we have to choose between the article and the training, the training is more interactive" led to the requirement that the training be interactive and not solely reading. Requirements were then divided into student and instructor functional requirements and content requirements (Table 36).

Table 36. Requirements devised from affinity diagram

Instructor Functional Requirements	Student Functional Requirements
Integrate training with course outcomes Must include instructions for instructors Simple integration with course materials Should not use additional tech Should be interactive Short in length	Incentive for completion Short in length Should be interactive/should not be solely reading
Content Requirements	
Should include the reasoning for completing the training Include expected positive outcomes from the training Leverage examples of what to do/what not to do Should include roadmap of the training process Examples and definitions should be relevant to everyday life	

Discussion

The results of the focus groups provide requirements for training as well as a deep dive into the perceptions of bias among instructors and students. Requirements for training content garnered from focus groups (e.g. the inclusion of examples guiding how one should evaluate a peer) echo the notion that bias reduction training should not only focus on awareness of bias, but also on actions that may be taken to correct the issue (Devine et al., 2012; Bezrukova et al., 2016). Multiple instructor functional requirements center on the workload shouldered by the instructors in implementing the training (e.g. simple integration, no new platforms). Therefore, reducing the workload associated with training implementation may increase training adoption. Instructors and students share the functional requirements of interactivity and short training length. Using interaction as opposed to text-based material is a technique proposed and used by researchers creating bias reduction training in other contexts (Devine et al., 2012; Olson and Harrell, 2020).

The requirements developed from the focus groups may be combined with the five challenges that should be addressed when creating bias reduction training (setting realistic

expectations for what training can accomplish, selecting proper goals, deciding how to manage discomfort, minimizing counter production, and demonstrating impact (Chapter 2, Carter, Onyeador, and Lewis, 2020)), the bias reduction method demonstrated by Devine and colleagues (2012), and the results of Chapter 4 to create a final set of training requirements and considerations (Table 37). These requirements inform the final design and deployment of training in the next chapter.

Table 37. Requirements and considerations for peer assessment bias reduction training

Instructor Functional Requirements	**Student Functional Requirements**
Integrate training with course outcomes Must include instructions for instructors Simple integration with course materials Should not use additional tech Should be interactive Short in length	Incentive for completion Short in length Should be interactive/should not be solely reading
Content Requirements	
Should include the reasoning for completing the training Include expected positive outcomes from the training Leverage examples of what to do/what not to do when assessing peers Should include roadmap of the training process Examples and definitions should be relevant to everyday life	
Five Considerations (Carter, et al.,2020)	
Set realistic expectations for what training can accomplish Select proper goals Decide how to manage discomfort Minimize counter production Demonstrate impact	
Devine Bias Reduction Method (Devine, et al., 2012)	
Bring attention to the issue of bias and its consequences Teach techniques to reduce biased actions Allow time for practice and learning	
Bias Considerations (Chapter 4)	
Gender Race/Ethnicity International student status Language	

Instructors noted in the focus group that while they had not been using formal peer assessment fairness training, they had taken steps to incorporate fairness into the peer assessment process. This is a similar finding to those discussed in Chapter 4, where instructors indicated in survey responses that they had infused fairness into peer assessments through informal trainings, discussions, and lecture material. Instructors already integrating fairness topics is a positive indication that the training itself can be "folded" into their peer assessment processes.

The focus groups also allowed the researchers a better understanding of how bias affects students and is recognized by instructors. Both students and instructors noticed bias due to gender, specifically in terms of task delegation and workload; female students were often delegated note takers or organizers and took on more work than male students. Students and instructors also talked about bias due to language. Students reported being unsure if a peer's performance was due to language barriers or a lack of effort, while instructors noted a general reluctance to comment on minority student performance. The content of the focus groups primarily centered on peer assessment feedback comments as opposed to peer assessment scores. As noted in the results in Chapter 4, students who are not native English speakers receive lower peer assessment scores overall. Synthesizing across these results, it is possible that while students are reluctant to comment critically on a peer's performance due to language, they are still willing to be critical in their numerical evaluation. A similar connection may be made due to race. Students report noticing bias due to race while instructors note that white students are unwilling to comment critically on the work of minority peers. Once again, students of color receive lower numerical peer evaluation scores than white students. This indicates that, as in the case of language, students reluctant to comment critically to a student of color are still willing to give them lower numerical scores.

Another prominent bias noticed by instructors and students was a general bias against the feedback process itself. Instructors noted that some students, particularly women and international students, are reluctant to give constructive feedback or lower assessment scores. Students note reasons for these actions including anxiety about their own evaluations (retaliation), not wanting to devalue the life circumstances of another student, or wanting to reward perceived effort.

The primary limitations of the focus groups are generalizability and the number of participants. All participants were from a Midwestern university with a primarily white student body, which could influence perceptions of racial bias by both students and instructors. This also impacts the generalizability of the results to more diverse higher education institutions. The low number of participants means that the results of the focus groups only reflect the experiences of a small number of individuals. While focus groups tend to involve small groups to encourage discussion, expanding the focus groups beyond three participants in each group could help to clarify some of the biases due to language and race.

Conclusion

Focus groups were conducted with instructors, students, and online learning professionals to learn more about classroom and peer assessment bias (RQ1) as well as gather feedback on the initial version of the training. Themes extracted from the focus groups were represented using an affinity diagram. From this diagram, functional requirements for students and instructors as well as content requirements were extracted. Requirements for both students and instructors focused on length, interactivity, and integration with existing course materials and technology. Content requirements centered on demonstrating the utility of the training, its expected outcomes, and the use of examples. These requirements form part of the answer to RQ3: *What are the requirements*

of and barriers related to implementing peer evaluation bias training in the classroom? The barriers discussed in the focus group, such as a need to add more technology, cut class content, or create additional course materials form the rest of the answer to RQ3.

CHAPTER 7: EVALUATION OF TRAINING TO IMPROVE PEER ASSESSMENT FAIRNESS

Research Objectives and Introduction

In Chapter 5, an initial training to address this bias was developed and evaluated to provide formative feedback to inform a design revision of the training. In this chapter, the final design and implementation of the training will be described. The training intervention was evaluated in a summative evaluation.

This work endeavors to combine work on bias reduction (e.g. Devine, et al.,) and appropriate assessment to create bias training specifically for the peer assessment classroom. The creation of the training began with the initial version of the training described in Chapter 5. This version was then iterated upon using the requirements discussed in Chapter 6. The specific objectives of this study were to iterate upon, implement, and evaluate the training in university classrooms. The iteration and implementation phases drew upon the requirements gathered, existing bias reduction methodologies, principles of effective feedback, Team-Based Learning literature, and best practices for training design. The evaluation of the training focused on student attitudes and perceptions of fairness at the beginning and end of the semester, as well as with each peer assessment.

Peer Assessment Fairness Training Iteration

As with the initial training, the final training was developed for deployment in Qualtrics in order to facilitate ease of implementation by instructors and eliminate the need for use of a class period. The training was divided into two parts: Part One – Giving Feedback and Part Two – Reducing Bias. Part One focused on the general process of giving useful and appropriate feedback in a peer evaluation while Part Two focused on recognizing and mitigating potentially

biased ratings and comments. The outlines of the initial and final trainings are shown in *Figure 22*. After learning information via videos or short readings, participants completed activities where this information was put to use. It has long been established that feedback is essential to effective learning (Bellon, Bellon, and Blank, 1991; Race, 2001; Yorke, 2002). Therefore, feedback was presented to students each time they submitted an answer to an activity.

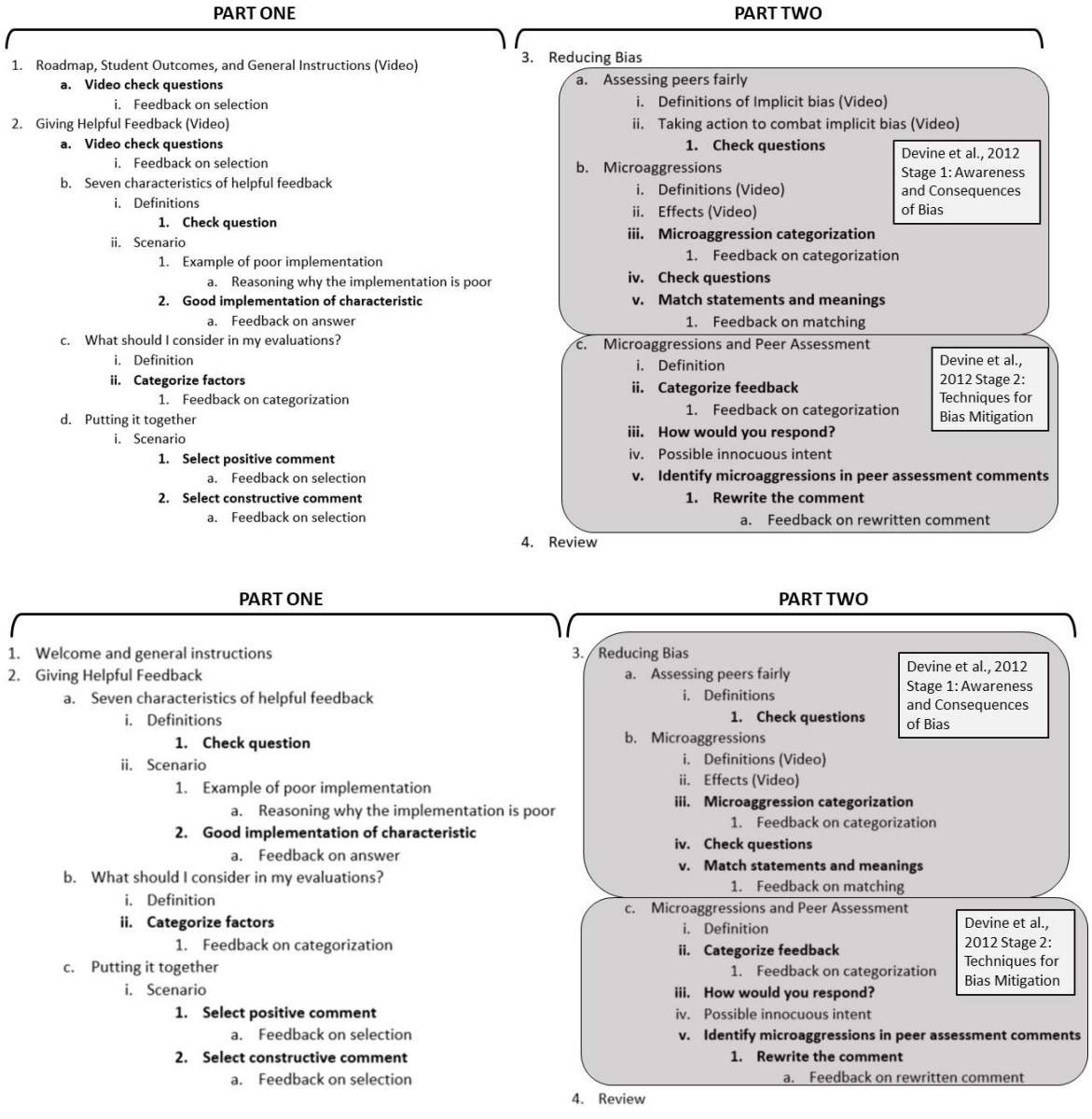

Figure 22. (Top) Outline of the final training; items in bold denote activities. (Bottom) Outline of the initial training; items in bold denote activities.

The content of the final training was updated based on feedback from the formative evaluation (see Chapter 5). In addition, the delivery mechanisms for much of the content were updated to better align with the requirements and considerations gathered in Chapter 6. The requirements and considerations are reviewed in Table 38. Requirements which necessitated changes to the training from the initial version are highlighted in bold in Table 38. Each bolded item will be discussed in this section, followed by a table demonstrating how the final training addressed all items in Table 38.

Table 38. Requirements for peer assessment fairness training. Bold items denote changes made between initial and final trainings

Instructor Functional Requirements	Student Functional Requirements
Integrate training with course outcomes **Must include instructions for instructors** **Simple integration with course materials** Should not use additional tech **Should be interactive** Short in length	Incentive for completion Short in length **Should be interactive/should not be solely reading**
Content Requirements	
Should include the reasoning for completing the training **Include expected positive outcomes from the training** Leverage examples of what to do/what not to do when assessing peers **Should include roadmap of the training process** Examples and definitions should be relevant to everyday life	
Five Considerations (Carter, et al.,2020)	
Set realistic expectations for what training can accomplish Select proper goals Decide how to manage discomfort Minimize counter production Demonstrate impact	
Devine Bias Reduction Method (Devine, et al., 2012)	
Bring attention to the issue of bias and its consequences Teach techniques to reduce biased actions Allow time for practice and learning	
Bias Considerations (Chapter 4)	
Gender Race/Ethnicity International student status Language	

Interactivity/Multimedia Integration

Requirements from both students and instructors included making the training interactive by using multimedia, as opposed to a set of readings. This interactive approach to training is well supported by the literature (e.g. Olson and Harrell, 2020) as it allows for intuitive use without increasing cognitive load (Schwan & Riempp, 2004).). The initial version of the training was interactive (see bolded items in Figure 22, bottom), however much of the information delivery was through text. The final version of the training utilizes short videos to deliver the majority of the content.

In Part 1, the training is introduced using a video made by the researcher. This video covers the purpose, benefits, roadmap, and general instructions (**Error! Reference source not found.**). The video was produced using Vyond, a commercially available video animation software, and is fully narrated and captioned for accessibility. Feedback is essential to effective learning (Bellon, Bellon, and Blank, 1991; Race, 2001; Yorke, 2002). Therefore, following the video is a set of check questions with feedback for the viewer to demonstrate they have watched the video. The script used in the video is found in Appendix A. The video itself may be viewed on YouTube by visiting the following web address: https://youtu.be/88vDZBNr0SE.

Part 1 of the training also covers the process for giving good feedback. In the initial training, characteristics of helpful feedback were presented via readings. For the final version, this information was presented in another narrated video created by the researcher (Figure 24). The video used the same characteristics of helpful feedback that had previously been presented in text (e.g. Michaelson and Schultheiss, 1988). The video also presented the items students should consider in their assessments (Onyia & Allen, 2012). Full descriptions of both the characteristics of helpful feedback and items students should consider may be found in Chapter

5. Information from the video was available as a reference in text or infographic form as well as in a link to view the video again during activities. The full video may be viewed on YouTube at the following web address: https://youtu.be/ry-SEb5HW-Y. The script used in the video may be found in Appendix B.

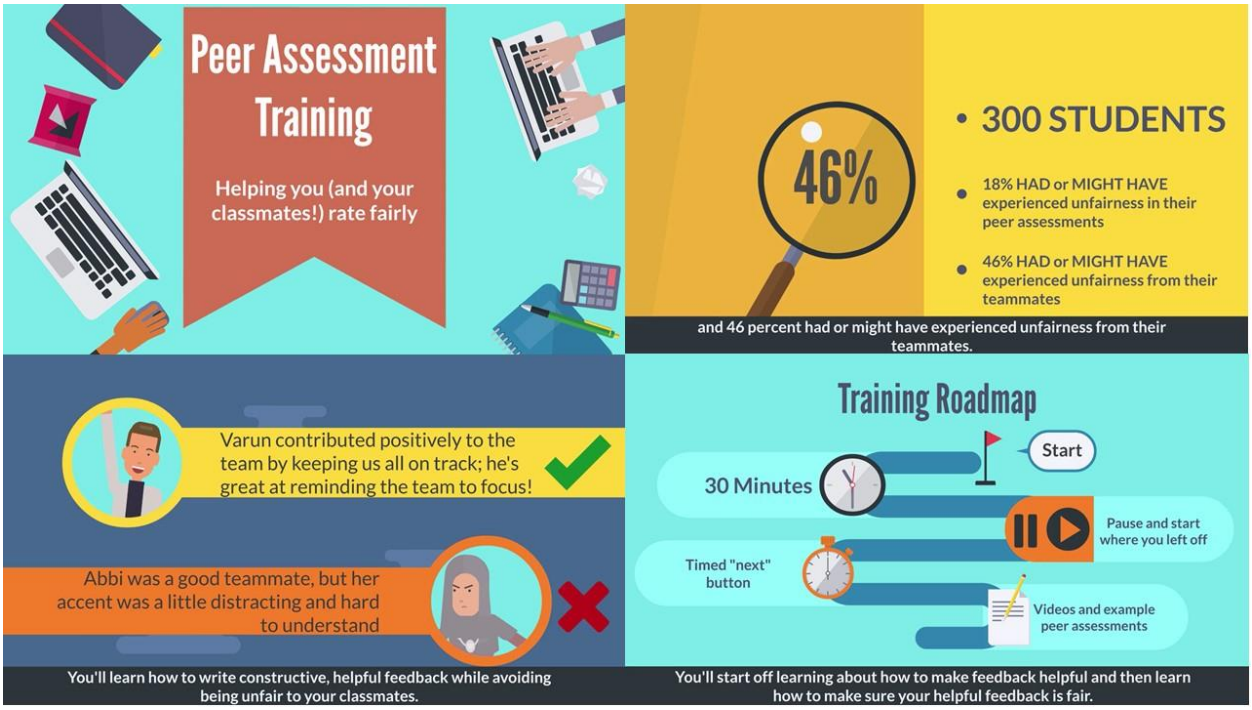

Figure 23. Screenshots from the introduction video

Part 2 of the training focused on how to reduce bias when assessing peers, and follows the Devine et al., (2012) three-stage model. The first stage of this model is awareness of bias and its consequences. In the initial version of the training, awareness of bias and its consequences were accomplished through readings. In the final training, bias and its consequences were covered using *Implicit Bias: Peanut Butter, Jelly, and Racism* (New York Times, 2019) and *High Heels, Violins, and a Warning* (Saleem Reshamwala, 2016). These videos were selected due to their short length, their framing of the topic of implicit bias in a manner to reduce but not eliminate discomfort, and their common use in implicit bias training materials. Short videos were

intentionally selected to keep the length of the training within requirements from students and instructors. Additionally, short educational videos have been shown to produce higher student-perceived retention, engagement, and focus than longer videos (Slemmons, Anyanwu, Hames, Grabski, Misna, Simkins, & Cook, 2018). The notion of reducing but not eliminating discomfort when learning about bias is drawn from recommendations for designing bias reduction training from Carter, et al., (2020).Information from the videos as well as links to rewatch the videos were available as a reference during activities.

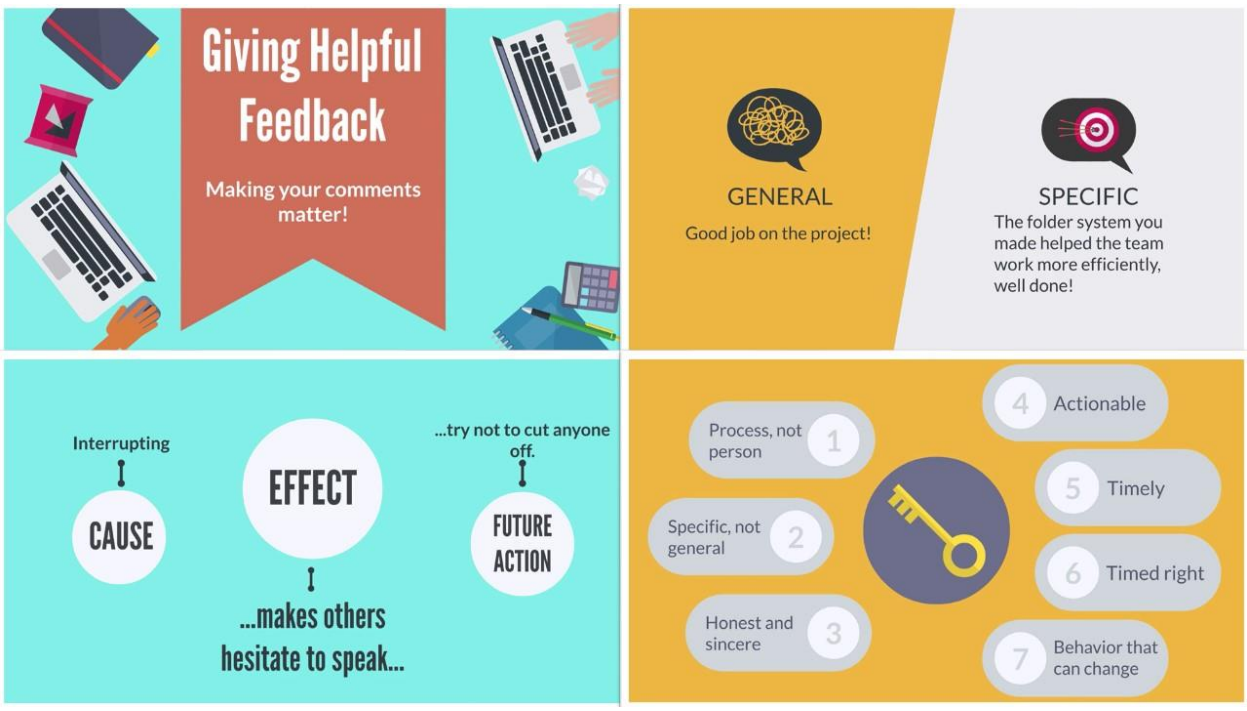

Figure 24. Screenshots from the helpful feedback instructional video

Instructions for Instructors/Simple Integration with Course Materials

The initial version of the training provided instructions for instructors via individual email. To make the instructions more readily accessible with the content of the training, a Canvas page was developed. This page contained instructions for using the training, overviews

of the training content, and the training materials itself (Figure 25). The need to create new course content is a barrier to adopting new methods or technology into the classroom (Beggs, 2012). Therefore, the Canvas page was also used to overcome that barrier by providing instructors with all needed materials.

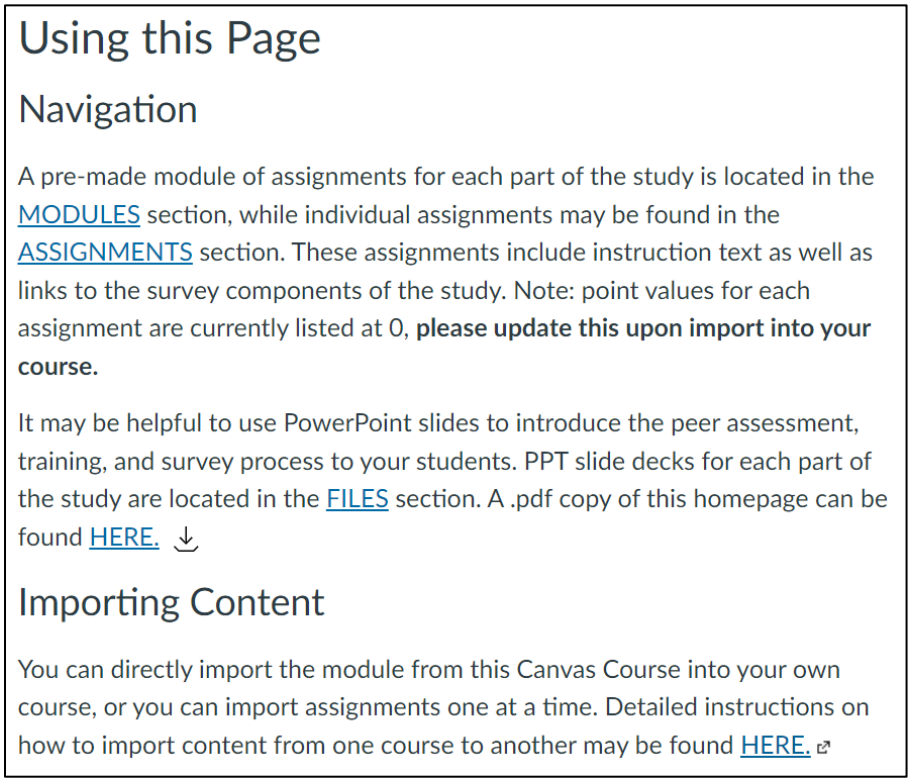

Figure 25. Snippet of Canvas page for instructors

To facilitate simple integration with course materials, all items for the training and its classroom evaluation were available for import as assignments from the training Canvas page to individual class Canvas pages (Figure 26). Instructions for using the page and importing the assignments were also provided, along with slide decks to use to introduce the content and the peer assessment process to the class.

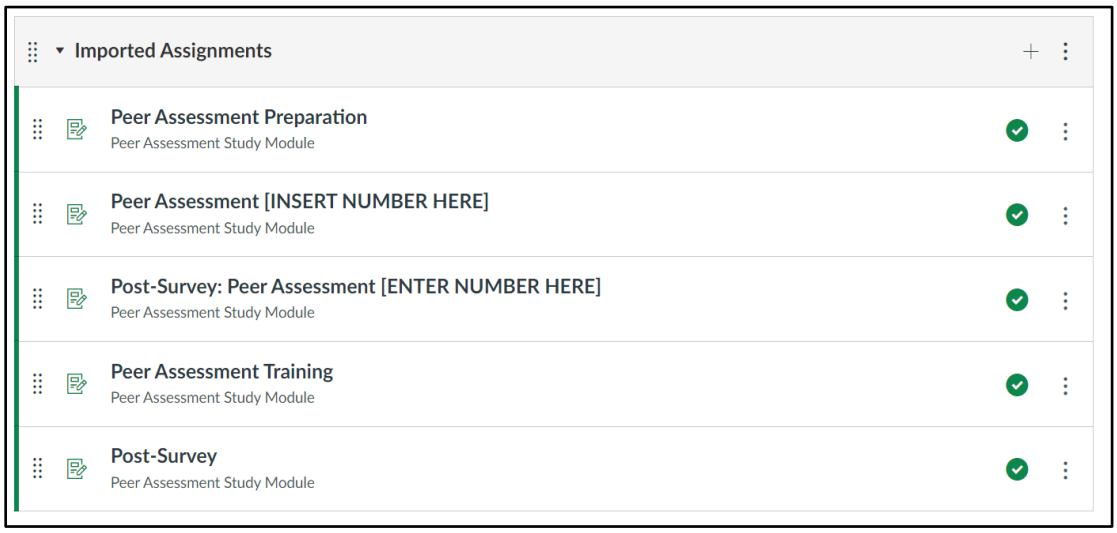

Figure 26. Snippet of assignments provided for instructors

Reasoning/Outcomes/Roadmap

Online learning professionals and instructors required that the training should have clear reasoning for its importance, expected outcomes, and a roadmap of what participants should expect. These items were presented to the students in the Introduction video noted above and in Figure 16, upper right. The importance of the training was framed using statistics from Chapter 4 (18% of students had or might have experienced unfairness in their peer assessments and 46% had experienced unfairness from their teammates). Students were informed that the expected outcome of the training was an understanding of how to give good, fair feedback, which would in turn increase the fairness and utility of assessments for them and their peers. Informing students of the expected training outcome correlates to setting learning objectives, which is a hallmark of effective teaching (Marzano, 2010). The training roadmap was presented as a visual road with markers noting important items (**Error! Reference source not found.**, lower right). Students were informed of the training length, that videos and activities included timers to prevent skipping, they could pause and continue on the same device as needed, and the type of

media they could expect (videos and activities). Training length and media type were disclosed to help set expectations and allow the students to plan their participation. Setting expectations in this way is recommended for effective classroom management (Marzano, 2010). Timers were included to ensure compliance. Students were allowed to pause and continue the training on the same device as this allows them control over their learning experience, and such control is associated with better learning outcomes (Schwan & Riempp, 2004).

Meeting Requirements and Considerations

All of the requirements and considerations developed in Chapter 6 were addressed in the final version of the training. The mapping between each requirement and how it was addressed is shown in Table 39.

Table 39. Mapping between each requirement and how it was addressed in the final training

Instructor Functional Requirements	Addressed
1. Integrate training with course outcomes **2. Must include instructions for instructors** **3. Simple integration with course materials** 4. Should not use additional tech **5. Should be interactive** **6.** Short in length	1. Training directly integrated with the peer assessment outcome in each course by requiring it as part of the peer assessment process **2. Instructions included in dedicated Canvas page for instructors** **3. Training and study materials integrated through Canvas** 4. Used the platform already in use, Canvas **5. Long readings replaced with videos and activities** **6.** ~30 minutes in length
Student Functional Requirements	Addressed
1. Incentive for completion 2. Short in length **3. Should be interactive/should not be solely reading**	1. Instructors included completion of the training in course grade 2. ~30 minutes **3. Long readings replaced with videos and activities**

Table 39 Continued	
Content Requirements	Addressed
1. Should include the reasoning for completing the training **2. Include expected positive outcomes from the training** 3. Leverage examples of what to do/what not to do when assessing peers **4. Should include roadmap of the training process** 5. Examples and definitions should be relevant to everyday life	1. **Reasoning and rationale for training included in introduction video** 2. **Expected outcomes included in introduction video** 3. Each skill introduced in the training includes examples of a poor vs. good implementation of the skill 4. **Roadmap included in introduction video** 5. Examples and definitions crafted to be relevant to the peer assessment scenarios students encounter
Five Considerations (Carter, et al., 2020)	Addressed
1. Set realistic expectations for what training can accomplish 2. Select proper goals 3. Decide how to manage discomfort 4. Minimize counter production 5. Demonstrate impact	1. Training should reduce perceptions of bias in peer assessment, but is not designed to address the larger issue of bias in all aspects of education. Training is only one part of creating fairer peer assessments in conjunction with actions instructors already take (e.g. diverse teams) 2. Training goals: create awareness of the issue of unfairness in peer assessments and what it looks like; teach strategies for giving fair peer assessments 3. Focus of biased actions taken off the individual and framed as an issue to which everyone may contribute 4. Training uses persuasive technique of presenting possible negative consequences of unfair peer assessments and strategies to avoid those consequences 5. Surveys on perceived fairness over time and analysis of peer assessment scores; interviews with students who took the training
Devine Bias Reduction Method (Devine, et al., 2012)	Addressed
1. Bring attention to the issue of bias and its consequences 2. Teach techniques to reduce biased actions 3. Allow time for practice and learning	1. Attention brought to the issue and its consequences through material in introduction video and videos in Part 2 of the training 2. Techniques for giving good feedback and reducing biased actions taught in Part 1 and Part 2 3. Training administered before second peer assessment to allow time for practice in subsequent assessments

Method

An Institutional Review Board (IRB) approved survey and interview study (20-055) of the use and effect of peer assessment bias training was conducted in five Iowa State University classrooms across the Spring 2021 and Fall 2021 semesters.

Participants

Participation in the study was at the class level. Instructors were recruited for the study via email. Participating instructors were then asked to add a statement to their syllabus explaining that the class was taking part in a study of peer assessment, and that the students may be asked to complete surveys and trainings related to the peer assessment process. Five classes took part in the study: IE 361 (S2021 and F2021), a junior-level Industrial Engineering class; CRP 561 (S2021), a graduate level Community and Regional Planning course; CRP 325 (F2021), a junior-level Community and Regional Planning course; and CRP 383 (F2021), a junior-level Community and Regional Planning course.

Across the five classes, 155 participants took part in the study. Fifty-five participants identified as women, 93 identified themselves as men, and seven preferred not to state a gender. One-hundred-eight participants identified as White, 15 participants identified as Asian, 12 participants identified as Black or African American, eight participants identified as Hispanic, four participants identified as another race or ethnicity, and eight participants chose not to answer. One-hundred-thirty-eight participants were native English speakers, while 17 were not. Participants also identified the college with which they were affiliated (Table 40).

Table 40. Participant counts by academic college (N = 155)

College	**Participant Count**
Agriculture and Life Sciences	6
Design	66
Engineering	83

Three sophomores, 76 juniors, 58 seniors, and 18 graduate students took part in the study. Sixteen participants reported being international students, while 139 were not. Participant-reported GPAs averaged 3.3 (ranged from 2.0-4.0).

Procedure

There were a total of nine activities in the project: a pre-survey, three self-assessments, three post-peer assessment surveys, the training, and a post-survey (*Figure 27*). All activities for the study were administered as assignments in Canvas which contained a link to the Qualtrics page for each item. Peer assessments were administered by the class instructor in Thinkspace or TEAMMATES (the peer assessment platform was at the discretion of the instructor). Participants were not divided into conditions (i.e. all students were assigned each activity as graded work), however individual participation in each activity varied.

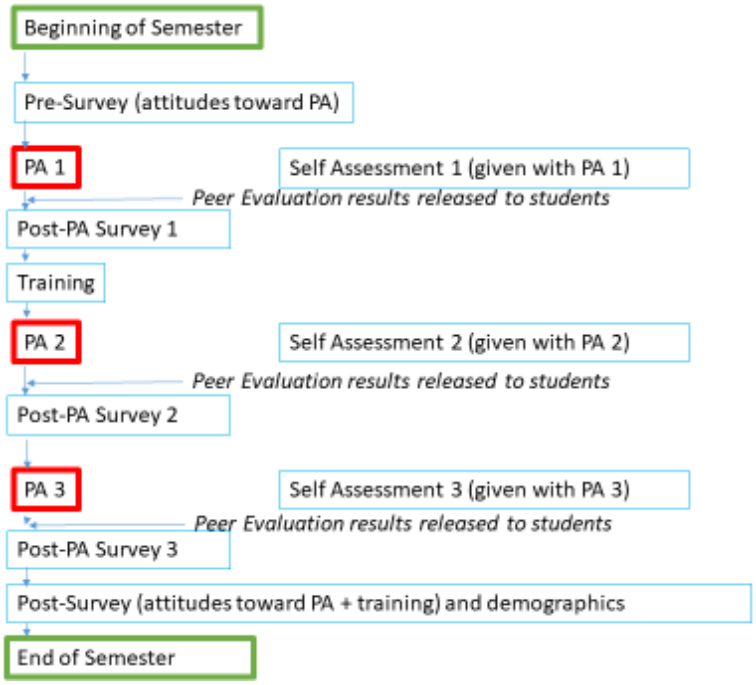

Figure 27. Timeline of activities in the study

Quasi-independent Variables

The study included two quasi-independent time variables: Study Time and PA Period. Study Time has two levels which correspond to the beginning of the semester (pre) and the end of the semester (post) (green boxes in Figure 15). PA Period has three levels; one for each peer assessment period (red boxes in Figure 15). Selected demographics were also compared. The variable "gender" was simplified into two categories: Men and Women due to lack of participants in the "other" and "prefer not to say" categories. The variable "race or ethnicity" was simplified into students of color (POC) and white.

Measures

The activities in the study may be broken down into measures by topic (Table 41). The measures used in the study did not change between the formative evaluation (Chapter 5) and the summative evaluation in this chapter.

Table 41. Measures used in the summative evaluation

Dependent Variable	Metric	Units	Assessment Frequency	Activity
Classroom team attitudes	4 survey questions	Likert scale 1-7	Study Time	Pre and post survey
PA attitudes and general fairness	8 survey questions	Likert scale 1-7, Likert scale 1-5 1 follow up free response	Study Time	Pre and Post survey
PA Fairness in specific class	5 survey questions	Likert scale 1-7, Likert scale 1-5 1 follow-up free response	Study Time (post-survey only)	Post survey
Reception of training	11 survey questions	Likert scale 1-5 3 follow up free response	Study Time (post-survey only)	Post survey
Peer Assessment	Michaelson method	0-15 points per team member 2 free response	PA Period	Peer Assessment
Self-assessment	Survey based on Michaelson method	0-15 points 2 free response	PA Period	Self Assessment
Fairness of PA Period	10 survey questions	Likert Scale 1-7, Likert scale 1-5 4 follow up free response	PA Period	Post Peer Assessment Survey
Demographics	10 survey questions	Multiple choice, free response	Study Time (post survey only)	Post survey

Classroom Team Attitudes

Questions on classroom team attitudes were used to determine whether there was a change in student perceptions of classroom teams from the beginning to the end of the semester. A full list of these questions may be found in Chapter 5.

PA attitudes and General Fairness

Questions on attitudes toward peer assessment and general peer assessment fairness were used to determine if students' perceptions of peer assessment and its fairness changed from the beginning to the end of the semester. A full list of these questions may be found in Chapter 5.

PA Fairness in Specific Class

Some questions on the fairness of peer assessments specifically in the class being studied were included only in the post survey. A full list of these questions may be found in Chapter 5.

Reception of Training

Questions on the reception of the training, its benefit, and methods of improvement were included only on the post survey. A full list of these questions may be found in Chapter 5.

Peer Assessments

Peer Assessments were administered using the Michaelson method. In this method, students assign a numerical score to their teammates based upon the extent to which they believe their teammates contributed to the team as a whole (Michaelsen, 2002). A full list of these questions may be found in Chapter 5.

Self-Assessments

The self-assessments, given with each peer assessment, were designed to be identical to the Michaelsen method peer assessments that the students completed. A full list of these questions may be found in Chapter 5.

Fairness of PA Period

The post-peer assessments, given after the students had the results of their peer assessment, were designed to gather the students' thoughts on the utility and fairness of the individual assessment. A full list of these questions may be found in Chapter 5.

Demographics

Demographics were assessed only at the end of the study as part of the post survey. This was done to prevent stereotype threat (Steele & Aronson, 1995). A full list of these questions may be found in Chapter 5.

Post Class Training Feedback

All students who had completed the training during the Spring 2021 semester were invited to participate in interviews about their experience during the Fall 2021 semester. Interviews were conducted via WebEx and lasted approximately 30 minutes. The researcher conducting the interviews took detailed notes on participant responses. Participants were compensated in the amount of $10 in cash. Four students participated in the interviews. Interview questions may be found in Table 42.

Table 42. Interview questions used with students in the summative evaluation

Interview Questions
What do you remember most about the training?
When you were completing the peer assessments, did you use information from the training? What did you use?
Do you think the training had an effect on the peer assessments you gave and received? What was the effect? How effective or ineffective was the training?
Were there differences in the peer assessments you gave and received before and after the training? What were the differences?
Have you used the information from the training in peer assessments outside of the class where you received the training? If so, what did you use and when?
What did you find most useful about the training?
What was least useful about the training or something that could be improved?
What is your gender?
How would you describe your race or ethnicity?
Are you an international student?

Data Analysis

Free response survey data was analyzed for overall counts and themes. Survey items which utilized Likert scales were analyzed using a one-way or multiway ANOVA for comparison over time (pre/post survey or peer assessment periods one, two, and three), and demographic variables. A significance level of $\alpha = .05$ was used throughout. Effect size was computed using Cohen's *d* (Cohen, 1988). The variable "gender" was simplified into two categories: Men and Women due to lack of participants in the "other" and "prefer not to say" categories. The variable for race or ethnicity was simplified into two categories: Person of Color and White. Participants were not divided into conditions; all students completed the training.

Results

Classroom Team Attitudes

The comparisons of team-based class attitudes by Study Time (pre and post) are summarized in Table 43.

Table 43. Comparison of team-based classroom attitudes by Study Time (pre/post). * indicates a significant result

Question		**Pre**		**Post**				
	N	*M*	*SD*	*M*	*SD*	*F*(1,N-1)	*p*	*d*
I like working in teams on class projects	155	5.1	1.3	5.4	1.3	3.54	.06	
I avoid classes that involve teamwork	155	2.6	1.3	3.0	1.6	7.38	.007*	0.27
When working on a team in a class or on a class project, I feel respected by my teammates	155	5.8	.9	5.9	.8	2.96	.09	
My classmates want to be on a team with me.	155	5.3	1.1	5.7	1.0	12.81	<.001*	0.38

Comparisons of team-based class attitudes are shown visually in Figure 28.

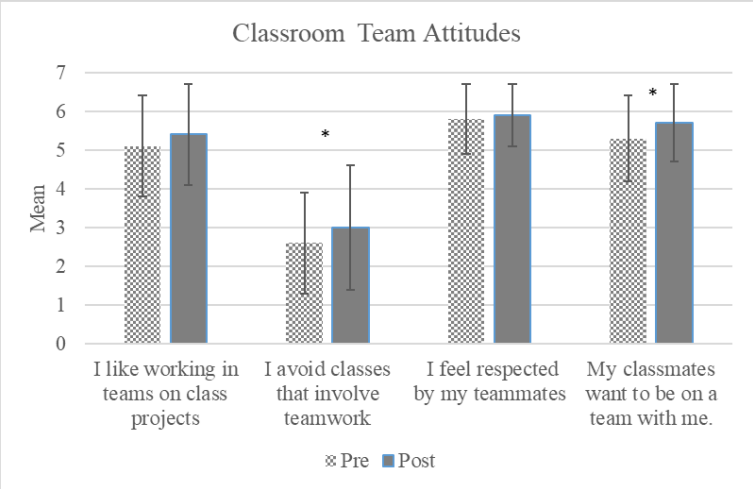

Figure 28. Classroom team attitudes by Study Time. Bars represent standard deviation; * indicates means which are significantly different.

The results for the main effect of gender on classroom team attitudes are summarized in Table 44. There were no significant interactions of gender and Study Time (pre/post).

Table 44. Effect of gender on classroom team attitudes

Question		**Men (N = 93)**		**Women (N = 55)**				
	N	*M*	*SD*	*M*	*SD*	*F*(1, N-1)	*p*	*d*
I like working in teams on class projects	148	5.4	1.1	5.4	1.0	0.92	.43	
I avoid classes that involve teamwork	148	3.1	1.0	2.9	1.0	0.82	.48	
When working on a team in a class or on a class project, I feel respected by my teammates	148	5.9	1.2	6.1	1.1	1.6	.20	
My classmates want to be on a team with me.	148	5.6	1.2	5.8	1.0	1.6	.20	

The results for the main effect of race on classroom team attitudes are summarized in Table 45. There were no interactions of race and Study Time (pre/post).

Table 45. Effect of race on classroom team attitudes

Question		**POC (N=39)**		**White (N=108)**				
	N	*M*	*SD*	*M*	*SD*	*F*(1, N-1)	*p*	*d*
I like working in teams on class projects	147	5.6	1.1	5.4	1.2	3.01	.051	
I avoid classes that involve teamwork	147	3.1	1.8	2.9	1.5	0.44	.64	
When working on a team in a class or on a class project, I feel respected by my teammates	147	6.0	.90	6.1	.82	2.65	.07	
My classmates want to be on a team with me.	147	5.7	.93	5.7	1.0	.13	.87	

The results for the main effect of English language status on classroom team attitudes are summarized in Table 46. There were no interactions of race and Study Time (pre/post).

Table 46. Effect of English language status on classroom team attitudes

Question		**English as First Language (N = 138)**		**English as Additional Language (N = 17)**				
	N	*M*	*SD*	*M*	*SD*	*F*(1, N-1)	*p*	*d*
I like working in teams on class projects	155	5.4	1.2	5.2	1.4	.20	.66	
I avoid classes that involve teamwork	155	2.8	1.5	4.0	2.0	8.44	.004*	0.68
When working on a team in a class or on a class project, I feel respected by my teammates	155	6.0	.92	5.8	1.1	.35	.55	
My classmates want to be on a team with me.	155	5.8	.92	5.2	1.2	5.00	.02*	0.56

The results for the main effect of international student status on classroom team attitudes are summarized in Table 47. There were no interactions of race and Study Time (pre/post).

Table 47. Effect of international student status on classroom team attitudes. * indicates a significant result

Question		International Students (N = 16)		Domestic Students (N = 139)				
	N	*M*	*SD*	*M*	*SD*	*F*(1, N-1)	*p*	*d*
I like working in teams on class projects	155	5.5	1.0	5.3	1.3	.28	.49	
I avoid classes that involve teamwork	155	4.1	2.0	2.9	1.5	8.57	.004*	0.68
When working on a team in a class or on a class project, I feel respected by my teammates	155	6.0	.78	5.9	.95	.067	.80	
My classmates want to be on a team with me.	155	5.5	1.1	5.7	.96	.79	.37	

PA attitudes and General Fairness

The comparison of peer assessment attitudes and general fairness by Study Time (pre and post) are summarized in Table 48.

Table 48. Peer assessment attitudes and general fairness by Study Time. * indicates a significant result

Question		Pre		Post				
	N	*M*	*SD*	*M*	*SD*	*F*(1,N-1)	*p*	*d*
Prior to this class, the peer assessments I received in other classes were fair	155	5.9	1.3	5.7	1.2	6.77	.012*	0.16
Prior to this class, the peer assessments I gave in other classes were fair	155	6.2	.65	5.9	1.0	13.07	<.001*	0.36
Prior to this class, how frequently have you felt unfairness in the peer assessments you have received?	155	1.6	.77	1.6	.80	0.003	.95	
I am confident I can assess my teammates fairly	155	6.0	.77	6.3	.71	13.57	<.001*	0.41
I am confident my teammates in this class can assess me fairly	155	6.0	.95	6.2	.72	6.77	.01*	0.24
Peer assessment is beneficial to my learning	155	5.2	1.4	5.2	1.5	0.034	.85	

The comparison of PA attitudes and general fairness are summarized visually in Figure 29.

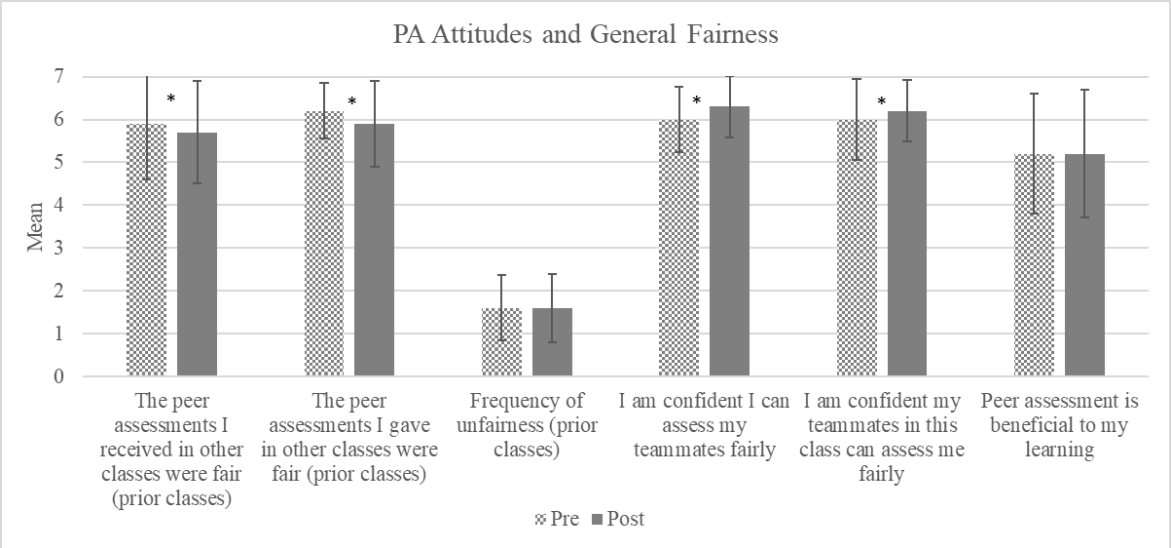

Figure 29. Peer assessment attitudes and fairness by Study Time. Bars represent standard deviation; * indicate means which are significantly different.

The results for the main effect of gender on peer assessment attitudes and fairness are summarized in Table 49. There were no interactions of gender and Study Time (pre/post).

Table 49. Effect of gender on peer assessment attitudes and fairness. * indicates a significant result

Question		**Men (N = 93)**		**Women (N = 55)**				
	N	*M*	*SD*	*M*	*SD*	*F*(1, N-1)	*p*	*d*
Prior to this class, the peer assessments I received in other classes were fair	148	5.7	1.1	5.6	1.2	0.37	.77	
Prior to this class, the peer assessments I gave in other classes were fair	148	6.1	1.2	5.8	1.3	1.4	.25	
Prior to this class, how frequently have you felt unfairness in the peer assessments you have received?	148	1.6	1.1	1.6	1.0	1.3	.28	
I am confident I can assess my teammates fairly	148	6.1	1.3	6.5	1.4	3.25	.02*	0.30
I am confident my teammates in this class can assess me fairly	148	6.2	1.5	6.3	1.2	.39	.76	
Peer assessment is beneficial to my learning	148	5.1	1.6	5.5	1.4	6.56	<.001*	0.27

The results for the main effect of race on peer assessment attitudes and fairness are summarized in Table 50. There were no interactions of race and Study Time (pre/post).

Table 50. Effect of race on peer assessment attitudes and fairness. * indicates a significant result

Question		**POC (N = 39)**		**White (N = 108)**				
	N	*M*	*SD*	*M*	*SD*	*F*(1, N-1)	*p*	*d*
Prior to this class, the peer assessments I received in other classes were fair	148	5.4	1.0	5.7	1.0	1.36	.26	
Prior to this class, the peer assessments I gave in other classes were fair	148	5.7	1.0	5.9	1.0	.86	.43	
Prior to this class, how frequently have you felt unfairness in the peer assessments you have received?	148	1.7	.89	1.5	1.0	.23	.80	
I am confident I can assess my teammates fairly	147	6.3	.78	6.2	.69	.67	.57	
I am confident my teammates in this class can assess me fairly	147	6.2	.62	6.2	.70	1.0	.37	
Peer assessment is beneficial to my learning	147	5.2	1.3	5.2	1.5	2.66	.07	

The results for the main effect of English speaker status on peer assessment attitudes and fairness are summarized in Table 51. There were no interactions of English speaker status and Study Time (pre/post).

Table 51. Effect of English speaker status on peer assessment attitudes and fairness. * indicates a significant result

Question		**English as First Language (N = 138)**		**English as Additional Language (N = 17)**				
	N	*M*	*SD*	*M*	*SD*	*F*(1, N-1)	*p*	*d*
Prior to this class, the peer assessments I received in other classes were fair	155	5.7	.99	5.2	1.1	4.20	.04*	0.48
Prior to this class, the peer assessments I gave in other classes were fair	155	5.9	1.0	5.7	.99	.68	.41	
Prior to this class, how frequently have you felt unfairness in the peer assessments you have received?	155	1.5	.71	2.0	1.3	5.15	.02*	0.48
I am confident I can assess my teammates fairly	155	6.2	.70	6.3	.77	.069	.79	
I am confident my teammates in this class can assess me fairly	155	6.2	.73	6.1	.70	.43	.51	
Peer assessment is beneficial to my learning	155	5.2	1.5	5.4	1.3	2.28	.60	

The results for the main effect of international student status on attitudes and fairness are summarized in Table 52. There were no interactions of international student status and Study Time (pre/post).

Table 52. Effect of international student status on peer assessment attitudes and fairness. * indicates a significant result

Question		**International Students (N = 16)**		**Domestic Students (N = 139)**				
	N	*M*	*SD*	*M*	*SD*	*F*(1, N-1)	*p*	*d*
Prior to this class, the peer assessments I received in other classes were fair	155	5.2	.97	5.7	1.0	2.70	.10	
Prior to this class, the peer assessments I gave in other classes were fair	155	5.8	1.2	5.9	1.0	.18	.67	
Prior to this class, how frequently have you felt unfairness in the peer assessments you have received?	155	2	1.2	1.6	.75	4.08	.04*	0.40
I am confident I can assess my teammates fairly	155	6.1	.86	6.3	..69	.34	.56	
I am confident my teammates in this class can assess me fairly	155	5.9	.73	6.2	.72	2.48	.12	
Peer assessment is beneficial to my learning	155	5.4	1.2	5.1	1.5	.51	.47	

PA Fairness in Specific Class

Items relating to the fairness of peer assessment in the class completing the study were evaluated using post survey data only. The results for the main effect of gender on PA fairness in a specific class are summarized in Table 53.

Table 53. Effect of gender on PA fairness in a specific class. * indicates a significant result

Question		**Men (N = 93)**		**Women (N = 55)**				
	N	*M*	*SD*	*M*	*SD*	*F*(1, N-1)	*p*	*d*
The peer assessments I received in this class were fair	148	5.8	1.1	6.3	1.1	3.00	.03*	0.45
In this class, how frequently have you felt unfairness in the peer assessments you have received?	148	1.4	.70	1.3	.52	3.43	.019*	0.16
The peer assessments I gave in this class were fair.	148	6.1	.83	6.4	.73	2.92	.04*	0.38

The results for the main effect of race on PA fairness in a specific class are summarized in Table 54.

Table 54. Effect of race on PA fairness in a specific class. * indicates a significant result

Question		**POC (N = 39)**		**White (N = 108)**				
	N	*M*	*SD*	*M*	*SD*	*F*(1, N-1)	*p*	*d*
The peer assessments I received in this class were fair	147	6.0	1.1	6.0	.88	.062	.96	
In this class, how frequently have you felt unfairness in the peer assessments you have received?	147	1.3	.61	1.3	.63	.055	.99	
The peer assessments I gave in this class were fair.	147	6.1	.81	6.2	.82	.84	.68	

The results for the main effect of English speaker status on PA fairness in a specific class are summarized in Table 55.

Table 55. Effect of English speaker status on PA fairness in a specific class. * indicates a specific result

Question		**English as First Language (N = 138)**		**English as Additional Language (N = 17)**				
	N	*M*	*SD*	*M*	*SD*	*F*(1, N-1)	*p*	*d*
The peer assessments I received in this class were fair	155	6.0	1.0	5.7	1.5	.76	.39	
In this class, how frequently have you felt unfairness in the peer assessments you have received?	155	1.4	.67	1.5	.80	.42	.52	
The peer assessments I gave in this class were fair.	155	6.2	.81	6.2	.75	.002	.96	

The results for the main effect of international student status on PA fairness in a specific class are summarized in Table 56.

Table 56. Effect of international student status on PA fairness in a specific class. * indicates a significant result

Question		**International Students (N = 16)**		**Domestic Students (N = 139)**				
	N	*M*	*SD*	*M*	*SD*	*F*(1, N-1)	*p*	*d*
The peer assessments I received in this class were fair	155	5.6	1.2	5.9	1.1	.92	.34	
In this class, how frequently have you felt unfairness in the peer assessments you have received?	155	1.4	.76	1.4	.69	.10	.75	
The peer assessments I gave in this class were fair.	155	6.3	.91	6.2	.80	.09	.76	

Part of the PA Attitudes and General Fairness measure included questions on fairness in peer assessment in classes taken prior to the class enrolled in the study and peer assessment fairness in the current class in which the participant received peer assessment training. There was a significant difference in perception of fairness of peer assessments received where assessments in the class with training (M = 5.9, SD = 1.1,) were perceived as fairer than the assessments in prior classes, (M = 5.6, SD = 1.0; $F(1, 154) = 5.9$, p = .01, d = 0.29). There was also a significant difference in perception of fairness of peer assessments given where assessments in the class with training (M = 6.2, SD = .81) were perceived as fairer than the assessments in prior classes (M = 5.9, SD = 1.0; $F(1, 154) = 10.0$, p = .002, d = 0.33). Additionally, students were more confident in the fair rating skills of their teammates from the class with training (M = 6.1, SD = .96) than teammates in prior classes (M = 5.7, SD = 1.1; $F(1, 154) = 11.8$, p <.001 , d = 0.39). These results are summarized visually in Figure 30

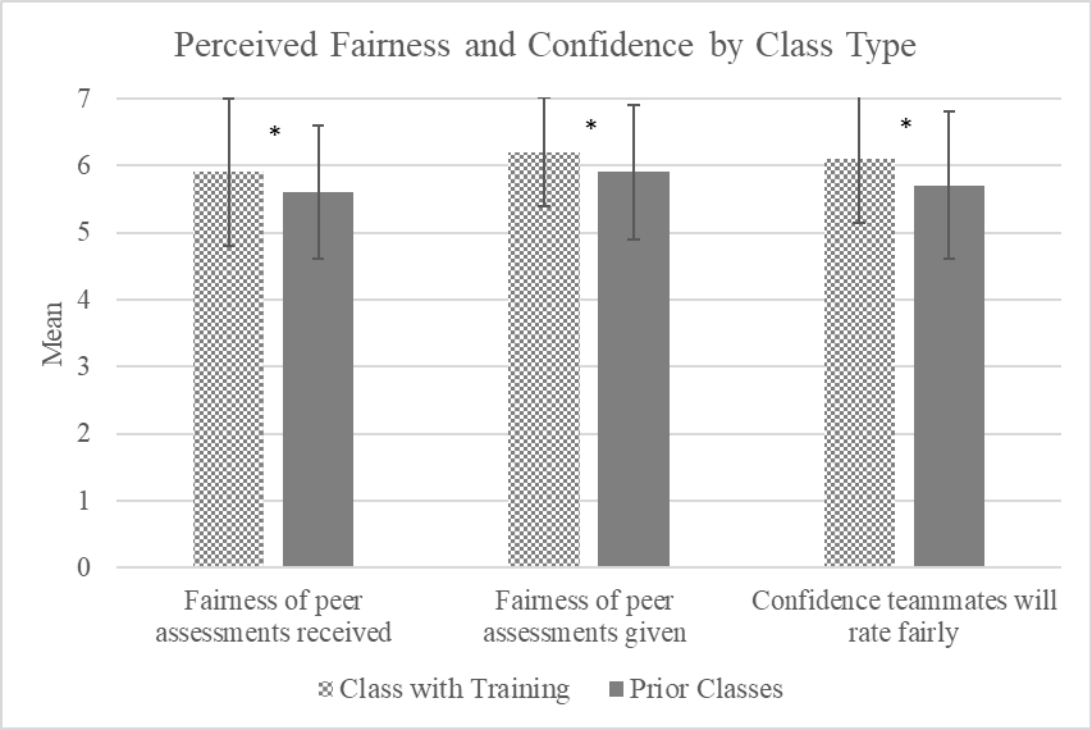

Figure 30. Fairness and confidence in ratings by class type. Bars indicate standard deviation; * indicate means which are significantly different

Reception of Training

The mean score for the benefit of the training was 3.51 (out of 5, *SD* = 1.2). Specific benefits of the training listed were: knowing what criteria to evaluate teammates against (31 mentions), examples of constructive comments (24 mentions), how to give actionable feedback (22 mentions), recognition of the importance of feedback from peers (11 mentions), learning what could be hurtful to a teammate and why (7 mentions), and instructions on how to craft a constructive comment (7 mentions). The mean score for using the information in the training when completing peer assessments was 3.9 (out of 5, *SD* = 1.1) and the mean score for feeling teammates used the information in the training was 3.6 (out of 5, *SD* = 1.1).

Study participants who were enrolled in additional classes which used peer assessment (but the additional classes were not part of the study) (N = 128) were asked to evaluate whether they used the training material in their other peer assessments (*M* = 3.5; *SD* = 1.3). Students indicated that the information they used from the training included: how to give thoughtfully worded comments (28 mentions), how to make constructive comments actionable (19 mentions), and avoiding personal bias (16 mentions). Suggestions for improving the training included: reducing length (30 mentions), making the training more personal (10 mentions), more interaction (7 mentions), and removing it altogether (21 mentions). The mean score for having discussed the training as a team was 2.97 (out of 5, *SD* = .80). Students who indicated their team had discussed the training mentioned discussing how it could be helpful (15 mentions), whether their peer assessments had been in line with the training material (9 mentions), and its length (9 mentions). Results for training reception are summarized graphically in Figure 31

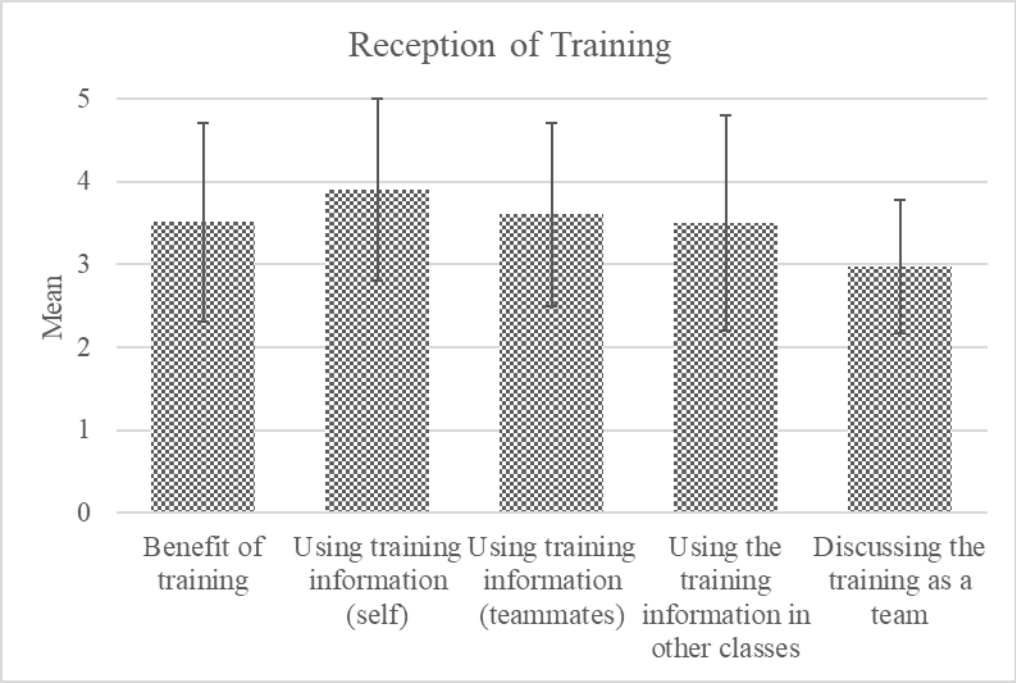

Figure 31. Reception of training. Bars indicate standard deviation

Retention

The reception and retention of the training was also explored through interviews. Of the four interview participants: two were men and two were women; one was Asian, one was Black, and two were white; two were international students and two were domestic students. Results in this section will be presented as phrases by interview question.

What do you remember most about the training?

- The videos were very helpful
- We were given examples and then had to apply what we learned
- The questions are clear and the language is clear and simple
- The videos told us what to look for in peer assessments
- A lot of it is common sense
- It seemed kind of dry
- People might be doing something a certain way that isn't their fault to make things unfair

When you were completing the peer assessments, did you use information from the training? What did you use?

- Yes, I used it. These trainings were really helpful for how to give constructive feedback, not just rating people
- I used it. How to communicate better in discussions and peer assessment, giving constructive feedback that isn't a microagression, not using mean terms
- I didn't use the training. I went off effort from the teammates.
- I used the training for writing my comments, but I didn't have much negative feedback for my teammates. I was in a pretty good group

Do you think the training had an effect on the peer assessments you gave and received? What was the effect? How effective or ineffective was the training?

- Yes. I feel usually the PAs were just about getting it done, the training helped us understand what different aspects there are to a teammate and how they can improve.
- Yes, it had an impact. When I did not know how to phrase constrictive criticism, I referenced the training for how to say it in a constructive way. Training was effective, the class that didn't have the training the comments were just there to get a grade or just because they had to. I had 2 PA classes at the same time. I got a mean comment in the non-training class, and noticed cruel comments in the PAs in the non-training class. In the PA training class, people tried to understand their team members. It was far better than the class where we didn't have training.
- I don't think so. They were all really busy, they all had to just go through and do it. We didn't take too much for the peer assessments. We wanted to give specifics for each member, but the training was something we just wanted to knock out. We hear this stuff all the time.

- Slightly positive, it didn't have a huge impact, but it made people be more thoughtful about the process. In the group, more nuanced feedback.

Were there differences in the peer assessments you gave and received before and after the training? What were the differences?

- Yes. The training helped me in giving constructive feedback. It's easy to say what's wrong, but to say it in a way to help them improve is hard. That's the change I could see for me and for the feedback I got. It made me understand the meaning of PA, not just about grading but focusing on improvement. The examples were helpful. It was kind of tiring to do, but it was very useful. The scenarios were helpful to me personally. Helped tune into what I really wanted to say in a more helpful less mean way.
- Something changed after training. People could just comment anything before training, we weren't thinking through it as much. Even I just commented for the sake of getting it done. After training, I thought of it as helpful feedback so they can change or keep it up. Training explains purpose of peer assessment. The training puts your brain into a place where you're doing it for the benefit of everyone. I knew the purpose of doing the PA, not just because I wanted my 5 points.
- I didn’t see any differences.
- The biggest thing was knowing how to make them more constructive and more helpful, more applicable and useful feedback more than “you’re a hard worker”, better actionable feedback. On the peer assessments I received, the feedback seemed more actionable and more channeled.

Have you used the information from the training in peer assessments outside of the class where you received the training? If so, what did you use and when?

- Yes. 1 PA class since, but it helps me in the long term with giving feedback in general.
- Yes. I used it being a TA. When grading, I try to use it to phrase the feedback to the students to make it useful but kind. I also use it in client presentations, for these we do peer feedback. I try to use it on the slide presentations too, to communicate what they could do better and they're doing well; I try to phrase like it is in the training.
- I've used the concepts we've talked about, heard this training used a lot. Not from this specific training, but when working with teams or groups, clarity is a big thing. I remember using the idea of specifics from the training in making rules of operation.
- No peer assessments outside the class. A couple assessments were a scale of 1-5 rate the teammate, but no written feedback at all since the class.

What did you find most useful about the training?

- Understanding how to give constructive feedback
- Training through examples and scenarios

- The phrasing examples.
- How you can get your message across, sometimes you don't know how. It can help overcome cultural differences in directness for what is acceptable and rude. How can I get my message across in a good way?
- Nothing
- How to make the feedback more effective, to make it help someone.
- How to give effective feedback that has a positive impact.

What was least useful about the training or something that could be improved on?

- A bit shorter.
- The videos were nice, I liked that I didn't have to read everything.
- Allow the videos to be played faster.
- The repetition was helpful to remember the material
- Use more illustrations and visuals.
- Instructors should be more aware of it, talk more about the WHY of peer assessments, and instructors should know the information in the training.
- The training should be given more broadly in more classes.
- I want to skip through it because there's nothing new.
- It's like an HR training, it's irritating.
- Break it down into parts or make it shorter
- The microaggression stuff doesn't really apply to the peer experience.

Peer Assessments and Self Assessments

Self-assessment scores (M = 10.5, SD = 1.2) were significantly higher than peer assessment scores (M = 9.4, SD = 1.4); $F(1, 882) = 108.2$, $p<.001$, d = 0.84. Peer assessment scores were significantly higher during the third assessment period (M = 10.2, SD = 1.1) than the first (M = 9.4, SD = .9), and second (M = 9.7, SD = 1.5); $F(2, 622) = 9.20$, $p <.001$, d = 0.50. Self-assessment scores were significantly higher during the second (M = 11.4, SD = 1.1) and third assessment periods (M = 11.7, SD = .9), than the first assessment period (M = 9.7, SD = 1.5); $F(2, 465) = 5.89$, p =.003, d = 1.29 . These results are summarized visually in Figure 32

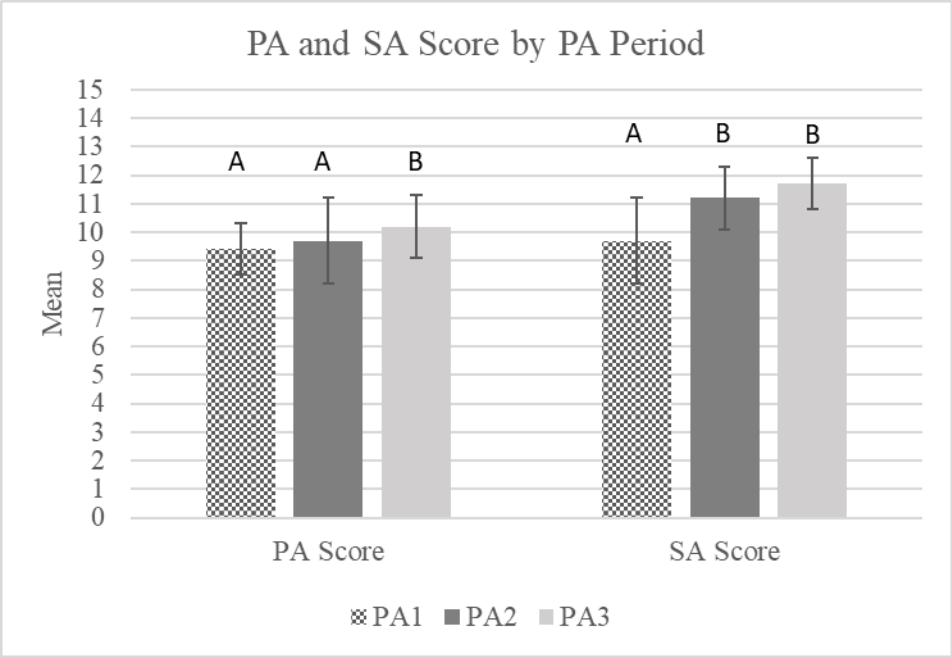

Figure 32. Peer assessment and self assessment score by PA period. Bars indicate standard deviation. Items not connected by the same letter are significantly different.

There were no significant differences in peer assessment score by gender, race or ethnicity, English speaker status, or international student status. There were no significant differences in self-assessment score by gender, race or ethnicity, English speaker status, or international student status.

Fairness of PA Period

There was a significant difference in perceived accuracy of the numeric feedback received where accuracy was higher in periods two (M = 6.4, SD = .7) and three (M = 6.5, SD = .8) than in assessment period one (M = 6.1, SD = 1.0) F(2, 358) 6.07, p = .003, d = 0.43. There were no significant differences in written accuracy of the peer assessments received for any of the assessment periods (P1: M = 6.4, SD = .9; P2: M = 6.3, SD = .8; P3: M = 6.2, SD = 1.0); F(2, 358) = .99, p = .38. There was a significant difference in perceived fairness of the peer assessment received where fairness was higher in period three (M = 6.5, SD = .6) than period one

(*M* = 6.2, *SD* = .7) and period two (6.3, SD = .8) *F*(2, 358) = 4.39, *p* = .013, *d* = . There were no significant differences in fairness of peer assessments given for any of the assessment periods (P1: *M* = 6.4, *SD* = .6; P2: *M* = 6.3, *SD* = .9; P3: *M* = 6.4, *SD* = .8); *F*(2, 358) = .69, *p* = .50. There was a significant difference in feedback being actionable where feedback significantly more actionable in periods one (*M* = 4.3, *SD* = .8) and two (*M* = 4.2, *SD* = 1.1) than in period three (*M* = 3.9, *SD* = 1.0) *F*(2, 358) = 5.73, *p* = .004, *d* = 0.44. There was a significant difference in intent to change behavior based on the peer assessment where intent was significantly higher in periods one (*M* = 4.5, *SD* = .6) and two (*M* = 4.3, *SD* = .8) than in period three (*M* = 4.0, *SD* = .9) *F*(2, 358) = 8.89, *p* <.001, *d* = 0.65 . These results are summarized graphically in Figure 33.

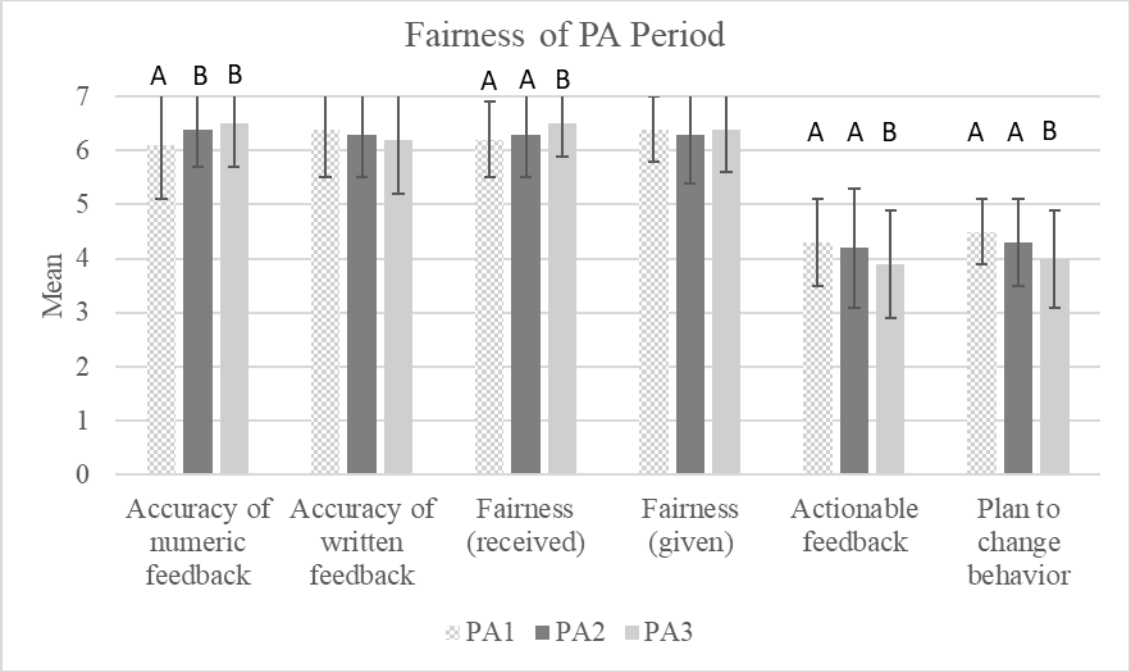

Figure 33. Fairness of PA period. Bars represent standard deviation; items not connected by the same letter are significantly different

There were no significant differences in perceived accuracy of the written feedback received between men (*M* = 6.2, *SD* = .9) and women (*M* = 6.3, *SD* = 1.0); *F*(1, 359) = .16, *p* =

.92. There were no significant differences in perceived accuracy of the numeric feedback between POC (M = 6.2, SD = .9) and White students (M = 5.8, SD = 1.1); F(1, 359) = 2.32, p = .13. Students whose first language was English found their written feedback to be significantly more accurate (M = 6.3, SD = .9) than students who learned English as an additional language (M = 5.8, SD = 1.2); F(1, 359) = 7.12, p = .008, d = 0.47. International students found their written feedback to be significantly less accurate (M = 5.8, SD = 1.0) than domestic students (M = 6.3, SD = .9); F(1, 359) = 4,11, p = .04, d = 0.53 . The interaction of PA Period and any demographic variable was not significant.

There were no significant differences in perceived fairness of the peer assessments received between men (M = 6.2, SD = .9) and women (M = 6.4, SD = .9); F(1, 359) = 1.26, p = .29. There were no significant differences in perceived fairness of the peer assessments between POC (M = 6.3, SD = .9) and White students (M = 6.0, SD = .9); F(1, 359) = 1.01, p = .31. Students whose first language was English found their peer assessment scores to be significantly more fair (M = 6.3, SD = .9) than students who learned English as an additional language (M = 5.8, SD = 1.2); F(1, 359) = 8.35, p = .004, d = 0.47. International students found their peer assessments to be significantly less fair (M = 5.8, SD = 1.0) than domestic students (M = 6.3, SD = .9); F(1, 359) = 7.04, p = .008, d = 0.53. The interaction of PA period and any demographic variable was not significant.

There were no significant differences in perceived fairness of the peer assessments given between men (M = 6.3, SD = .9) and women (M = 6.5, SD = .8); F(1, 359) = 1.85, p = .11. There were no significant differences in perceived fairness of the peer assessments given between POC (M = 6.4, SD = 1.0) and White students (M = 6.5, SD = .9); F(1, 359) = .21, p = .64. Students whose first language was English found their peer assessment scores given to be significantly

more fair (M = 6.4, SD = .9) than students who learned English as an additional language (M = 6.0, SD = 1.1); $F(1, 359) = 4.39$, $p = .03$, $d = 0.40$. There were no significant differences in perceived fairness of the peer assessments given between international (M = 6.4, SD = 1.0) and domestic students (M = 6.2, SD = .8); $F(1, 359) = 1.14$, $p = .29$. The interaction of PA Period and any demographic variable was not significant.

There were no significant differences in finding their feedback actionable for women (M = 4.2, SD = .8) and men (M = 4.0, SD = .8); $F(1, 359) = 1.20$, $p = .32$. There were no significant differences in finding their feedback actionable for POC (M = 4.1, SD = .8) and white students (M = 4.9, SD = .7); $F(1, 359) = .54$, $p = .46$. There were no significant differences in finding their feedback actionable for students whose first language is English (M = 4.2, SD = .8) and students who learned English as an additional language (M = 4.1, SD = .8); $F(1, 359) = .59$, $p = .44$. There were no significant differences in finding their feedback actionable for international students (M = 4.2, SD = .8) and domestic students (M = 4.1, SD = 1.0); $F(1, 359) = .64$, $p = .42$. The interaction of PA Period and any demographic variable was not significant.

There were no significant differences in likelihood of changing behavior based on feedback for women (M = 4.9, SD = 1.3) and men (M = 4.6, SD = 1.2); $F(1, 359) = .97$, $p = .18$. There were no significant differences likelihood of changing a behavior for POC (M = 4.6, SD = .8) and white students (M = 4.8, SD = .8); $F(1, 359) = .15$, $p = .69$. There were no significant differences in likelihood of changing a behavior for students whose first language is English (M = 4.6, SD = .8) and students who learned English as an additional language (M = 4.6, SD = .8); $F(1, 359) = .021$, $p = .88$. There were no significant differences in likelihood of changing a behavior for international students (M = 4.7, SD = .7) and domestic students (M = 4.6, SD = .9);

$F(1, 359) = .49$, $p = .48$. The interaction of PA Period and any demographic variable was not significant.

Discussion

The initial evaluation of the training showed no significant changes in confidence, fairness, or perceived accuracy. Based on the feedback from the initial evaluation, the training was modified to address issues in interactivity and content delivery. The evaluation of the second iteration of training material showed significant positive changes in confidence, fairness, and perceived accuracy of feedback. Students also found their peer assessments in the class where they received training to be fairer than those received in previous classes. Further, participants were more confident in their trained teammates rating skills than in the rating skills of potential future teammates. These results may indicate that the training is bolstering students' confidence in their fair peer assessment skills and their perceptions of peer assessment fairness overall. Furthermore, in both the initial and final training evaluations, data from the pre- and post-surveys indicate an increased awareness of biased peer assessments, which corresponds to the first step in the Devine model of bias mitigation (Devine, et al., 2012).

In interviews, some students echoed and expanded upon this confidence stating that after training, the quality of peer assessment they gave and received increased. One student compared the comments received in the class with training to another class without training and found the comments in the un-trained class to be "cruel" while the comments in the trained class were "far better". Students also discussed using the training material for feedback in other peer assessments, as a teaching assistant, and as general good communication practices.

The training itself was generally well-received, with a mean score for its benefit of 3.5 (out of 5), which is a marked improvement from the score of 2.7 received by the initial version of

the training (Chapter 5). Additionally, the mean response for using the material from the training was 3.9 (out of 5). Taken together these results suggest that, while it could be improved, the material in the training is being used as intended in peer assessments.

In both the post survey (at the end of the study) and interviews (the semester after the study concluded), students mentioned that the training specifically helped with writing actionable feedback. However, in the post-peer assessments, students rated their feedback as being significantly less actionable in the second and third PA periods than the first. This may be analogous to the concept of diminishing returns – students receive feedback and, ideally, take corrective action (Castaño-Muñoz, Sancho-Vinuesa, & Duart, 2013). Therefore there may be less for them to take action on as they progress through the class. This indicates that giving the training before the first peer assessment could increase its benefit as that timing would allow students to apply their knowledge of actionable feedback where it can be most useful.

There are many results for perceived fairness in this smaller study which does not echo those described in the large-scaled analysis of five years' worth of peer evaluation data (see Chapter 4). For example, while in Chapter 4 women reported lower levels of perceived fairness via the student survey and received lower peer assessment scores overall in the large scale data analysis, in this study there were no significant differences by gender for perceived fairness in the pre/post surveys or in the peer assessment data itself. This is similar to the mixed findings discussed in Chapters 1 and 2, where indications of bias are inconsistent. However, like many of the studies reporting mixed results (e.g. Pajares & Johnson, 1996; Kaufman, Felder, & Fuller, 2000; Bryan, Krych, Carmichael, Viggiano, & Pawlina, 2005; Sherrard, Raafat, & Weaver, 1994; May & Gueldenzoph, 2006; Tucker, 2014), and in contrast to the results of Chapter 4, this study employed a small number of classrooms. This finding highlights the importance of

examining a large body of data for evidence of bias, as on the small-scale, individual student experiences differ. The findings in this study specifically, especially for students of color, are hindered by low numbers and the demographics of the institution in which it was conducted. Different groups (e.g. Asian students, Hispanic students) may have very different experiences with peer assessment, but were analyzed as a whole.

However, in some areas there are very consistent findings regardless of number of participants. For general fairness and fairness of PA period in this study as well as the student survey and large scale data analysis in Chapter 4, students who learn English as an additional language report lower fairness and receive lower peer assessment scores. This finding is similar to those reported by Langan, et al., (2005), which could indicate that the experience of these students is more consistently biased than other students.

Self-assessment scores were significantly higher than peer assessment scores in all assessment periods. This phenomenon is well-researched and generally expected (Golightly, 2021; Evans, Leeson, & Petrie, 2007; Campbell, Mothersbaugh, Brammer, & Taylor, 2001). However, it has also been reported that women consistently under-rate themselves in comparison with their male peers (Rees, 2003; Lind, Rekkas, Bui, Lam, Beierle, & Copeland, 2002; Das, 1998). This was not the case in this study where there were no significant differences in self-assessment scores by gender.

The study was limited by the demographics of the institution in which it was conducted, the limited number of classrooms where the training was deployed, the length of the training, and the characteristics of the training developer. The institution where the training was evaluated has a heavily white and domestic student population. This could limit the breadth of experiences captured in surveys as well as in reception to the training, which in turn can limit the

generalizability of the results to more diverse student bodies. The training was deployed in five classrooms, however the classrooms were all within two departments (Industrial Engineering and Community and Regional Planning). Further testing the training in different academic departments (e.g. liberal arts departments) could bolster the generalizability of the results. The length and timing of the training was crafted so as to not add greatly to student workload or necessitate removal of class material. However, one-time trainings are often not as effective as trainings over time (Carter, Onyeador, and Lewis, 2020). This could be addressed by offering the training in parts or adding follow-up activities into course content to reinforce training material. Caution should be taken, though, as there is a balance between the "most perfect" training and the most feasible training. Finally, the characteristics of the training developer could impact the generalizability of the training to all student populations. It is well established that designers design in their "own image" (Moss & Gunn, 2007; Moss, 2003). Therefore, as the training designer was a white woman, the content could be skewed to align more with that perspective. Creating a diverse design team could address this issue.

Conclusion

In this chapter, the final version training to address peer assessment fairness was described and evaluated. While both bias and assessment have been targets of previous training endeavors, this work combines lessons learned from the formative training evaluation and requirements gathering studies to design effective peer assessment training. The results of this evaluation of the training show evidence of improved perceptions of peer assessment fairness and student confidence in their and their classmates' rating abilities after completing training. No evidence of bias was found in peer assessment scores associated with the study, which is in contrast to the results of the large scale data analysis described in Chapter 4. However, a key

difference that could explain this discrepancy is the much smaller number of evaluations examined in the current study. These results indicate that the training could improve peer assessment outcomes for students while bolstering instructors' confidence in their use of peer assessment.

CHAPTER 8: CONCLUSIONS AND CONTRIBUTIONS

Review of Problem

The purpose of this project is to understand the issue of bias in peer assessment and design, implement, and evaluate training to mitigate or reduce these biases. The employment of small group, active learning strategies in classroom environments has been shown to increase student achievement, attendance, engagement, and lead to better overall learning outcomes (Michaelson, Knight, & Fing, 2004; Michaelson & Sweet, 2011; Allen, Copeland, Franks, Karimi, McCollum, Riese, & Lin, 2013). Because of these outcomes, team-based pedagogies and cooperative learning practices have been incorporated in higher education to improve the classroom engagement of underrepresented students. Indeed, research shows that learning in teams positively affects objective outcomes (such as exam scores) for minority students (Slavin & Oickle, 1981; Springer, Stanne, & Donovan, 1999). In engineering classrooms, active learning strategies have specifically been recommended by organizations such as ABET (Lima, Andersson, and Saalman, 2017). Researchers have validated the use of learning teams in engineering classrooms as better approaches to enhance acquisition of material (Freeman, et al., 2014). In many group and active learning classrooms, peer assessments are used to ensure individual accountability. By understanding that their contributions to the team will be assessed and potentially included in a grade, students are less likely to engage in "social loafing" (failure to participate). Therefore, they are more motivated to participate and contribute (Cestone, Levine, and Lane, 2008). Effective peer assessments also contribute to improved team functioning and a sense of belonging among teammates (Brown et al, 2021).

While the learning outcomes of these pedagogies are particularly positive for underrepresented or minority students, their experiences with the associated peer assessments are often not. Both students and instructors have expressed concerns about the fairness of teams and their associated peer assessments, especially due to bias (Magin & Helmore, 2001; Samuel, 2004; Dancer & Kamvounias, 2005; Aryadoust, 2016). Research, including Chapter 4 of this dissertation, has shown that the experiences of women and students of color in these classrooms differ from those of their peers in terms of assessment (e.g. Wayland et al., 2014, Chapter 4). Because of these concerns, interest in designing a fairer peer assessment process has increased. By understanding where and how bias occurs in peer assessment, training can be designed to directly target problem areas. This training could then be used to improve the fairness of assessments, which in turn could ensure that the positive outcomes associated with learning teams are shared among all students.

Review of Approach

The approach to addressing this problem was divided into three levels which correspond to different parts of the work. The goal of Phase 1 was to explore the problem of bias in peer assessment. This was accomplished through literature review, surveys of students and instructors on perceptions of peer assessment bias, and an analysis of over twenty thousand peer assessment ratings given and received within the Thinkspace peer assessment platform. In Phase 2, the bias mitigation training began to take shape. Using the lessons learned from an initial in-class pilot of bias mitigation methods as well as the results of the studies in Phase 1, the first version of the online training was developed. This training was then deployed in a limited number of classrooms and used as a starting point for gathering requirements from stakeholders. In Phase 3,

the training was further refined according to the requirements developed in Phase 2. The refined training was then evaluated and prepared for further use.

Findings

The research questions associated with this approach are as follows:

R1. How do students and instructors perceive bias in peer evaluations?

The first research question focuses on attitudes and perceptions of bias in peer assessment, and how these may differ with perspective (student versus instructor). Both students and instructors perceived evidence of biased peer assessments. Bias in their classroom peer evaluations was noticed by 47.3% of instructors. These biases were attributed to gender, race, interpersonal relationships, language, and gaming the peer assessment system. Classroom bias when working on teams was felt by 27.3% of students and might have been felt by 18.5% of students. These biases were due to gender, race, age, and interpersonal relationships. Further, peer assessment bias was perceived by 9.2% of students and might have been perceived by 9.2% of students. Similarly to classroom bias, these peer assessment biases were commonly due to gender, personality, and interpersonal relationships. When asked to rate the fairness of their peer assessments, women and students for whom English was an additional language reported significantly lower fairness than men and native English speakers.

R2. What evidence of bias is present in peer evaluation data?

The second research question looks to examine the occurrence of bias empirically through analysis of a large body of peer assessment data. Together, R1 and R2 lead to a better understanding of where bias happens, to whom bias is directed, and its prevalence. This understanding of the types of bias most prevalent in peer evaluations is important when crafting relevant training materials. Previous work has documented evidence of peer assessment bias in

individual classes, however the aim of these research questions is to look for bias across varying disciplines in order to create more generalizable training.

In this work, we found evidence of bias in a dataset of over 20,000 peer assessment ratings. Males received significantly higher peer assessment scores than females. International students and students for whom English was an additional language received significantly lower peer assessment scores than domestic US students and native English speakers. Students of color received significantly lower peer assessment ratings than white students. For women, English language learners, and international students, peer assessment scores and GPA move in opposite directions. Conversely, for students of color, peer assessment scores and GPA move in similar directions. While students of color did have lower GPAs than white students, there are many factors affecting the GPA of racially marginalized students (e.g. being a working learner) that were not analyzed. There is limited work showing a positive correlation between GPA and peer assessment score (Al Mortadi, et al., 2020). However, due to the limited scope of the Al Mortadi, et al. (2020) study, GPA is not suggested as an exact correlation to the team skills measured by peer assessment. Therefore, it is difficult to determine the exact cause of the lower peer assessment ratings received by students of color. Lower peer assessment scores could be due to factors such as bias (e.g., Thondlana & Belluigi, 2017), student team performance and grading (e.g, .ONeill, Boyce, & McLarnon, 2020), GPA (e.g. Al Mortadi et al., 2020), or a combination of multiple factors.

R3. What are the requirements of and barriers related to implementing peer evaluation bias training in the classroom?

The third research question addresses the functional requirements of bias mitigation training as well as potential barriers to training adoption. Following the user-centered design

process, the identification of requirements represents the first step in creating a training which meets the needs of its users. This research question also seeks to understand barriers to participation in training. As willingness to adopt a new method or intervention into the classroom can vary widely, knowledge of potential barriers to adoption and how to overcome them may improve instructors' willingness to use the intervention in their classroom (Monahan, McDaniel, George, & Weist, 2014).

In this work, we determined the requirements for peer assessment fairness training. These requirements were drawn from the results of focus groups with students, instructors, and online learning professionals as well as the literature. Students and instructors share multiple requirements including training that is short in length and interactive. Instructors have additional functional requirements, such as not using additional technology, to facilitate ease of integration into their classroom. Content requirements developed from all three groups included the use of real-life examples, explanation of the utility of the training, and discussion of the expected outcomes.

Barriers to implementing peer assessment fairness training were determined from the results of the focus groups. Barriers include: time to add training to course, needing to cut course content to add training, learning and integrating new technology or platforms to host and deploy the training, and needing to create new course materials surrounding the training. These barriers are similar to reported barriers to adopting new instructional technology in the classroom (e.g. time to learn new technology, development of new materials) (Beggs, 2012).

R4. Does bias mitigation training positively impact student perceptions of peer assessment fairness?

The fourth research question focuses on the efficacy of the training. Students often view the peer assessment process as unfair (e.g. Wayland et al., 2014) due to the potential for biased ratings or perceived lack of qualification of their peers. Evaluation of the proposed intervention aims to improve perception of peer assessment fairness, which in turn may have positive impacts on student willingness to fully participate in peer evaluation.

In this work, we found evidence that the training increases student perceptions of peer assessment fairness. Student-reported ratings of peer assessment fairness was significantly higher in peer assessments conducted after receiving the final training. Additionally, student confidence in their ability to rate fairly was significantly higher after receiving training. Students also found the assessments in their classes with training significantly fairer than their classes without the training.

Contributions and Future Work

The contributions of the work consist of a broad understanding of the issue of bias in peer assessment as well as a training to increase peer assessment fairness. Specially, evidence of bias has been found from the student perspective, the instructor perspective, and peer assessment data itself. While evidence of bias has been found previously, the results have been mixed due to low sample size and the inclusion of only one class or department. The large numbers of participants across academic departments in the studies described here bias address these concerns. The training developed to address this issue resulted in better perceptions of fairness and increased student confidence in their rating ability.

In order to make group based active learning strategies effective, instructors must encourage individual accountability. Peer assessment is a useful method for ensuring this accountability. However, if the peer assessment system is flawed and unfair due to bias, this puts

the gains from using active learning at risk. This is particularly true for minority groups and women. The training was designed to be appropriate across disciplines in higher education in order to maximize its utility. Fairer peer assessment provides enhanced access to the benefits of active and team learning to a broader range of students, potentially higher retention of women and minority students (for which there is a great need in engineering), and improved ABET measurement of student outcome 5.

Further questions in this area remain unanswered by this project. As noted previously, there is great difficulty in assessing "friendship" among team members, but this has long been posited as a contributing factor in biased peer assessments. Developing a large-scale measurement strategy for interpersonal relationships among team members could be a pertinent first step in understanding the occurrence of friendship bias. Additionally, this project only tracks students through one semester (approximately 15 weeks). However, other bias interventions have shown lasting positive effects at and beyond six months post-intervention (Devine, et al., 2012). It may be beneficial for future work to track students' attitudes after they exit the class in which they completed the bias mitigation training to understand the training's longevity.

REFERENCES

ABET. (2020). Criteria for Accrediting Engineering Technology Programs, 2020 – 2021. Retrieved January 15, 2021, from https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-technology-programs-2020-2021/

Al Mortadi, N., Al-Houry, S. S., Alzoubi, K. H., & Khabour, O. F. (2020). Effectiveness of Peer Evaluation in Learning Process: A Case from Dental Technology Students. The Open Dentistry Journal, 14(1).

Allen, R. E., Copeland, J., Franks, A. S., Karimi, R., McCollum, M., Riese, D. J., & Lin, A. Y. (2013). Team-based learning in US colleges and schools of pharmacy. *American journal of pharmaceutical education*, *77*(6).

Apfelbaum, E. P., Sommers, S. R., & Norton, M. I. (2008). Seeing race and seeming racist? Evaluating strategic colorblindness in social interaction. *Journal of personality and social psychology*, *95*(4), 918.

Archer, J. (1992). Sex bias in evaluations at college and work. *The Psychologist: Bulletin of the British Psychological Society*, *5*(5), 200-204.

Arcuri, L., & Boca, S. (1996). Pregiudizio e affiliazione politica: destra e sinistra di fronte all'immigrazione dal terzo mondo. *Psicologia e politica*, 241-273.

Aryadoust, V. (2016). Gender and academic major bias in peer assessment of oral presentations. *Language Assessment Quarterly*, *13*(1), 1-24.

Baker, D. F. (2008). Peer assessment in small groups: A comparison of methods. *Journal of Management Education*, *32*(2), 183-209.

Barak, M., & Rafaeli, S. (2004). On-line question-posing and peer-assessment as means for web-based knowledge sharing in learning. *International Journal of Human-Computer Studies*, *61*(1), 84-103.

Barak, M., Watted, A., & Haick, H. (2016). Motivation to learn in massive open online courses: Examining aspects of language and social engagement. *Computers & Education*, *94*, 49-60.

Bartlett, T. (2017). Can we really measure implicit bias? Maybe not. *The chronicle of higher education*, *63*(21), B6-B7.

Beggs, T. A. (2012). Influences and Barriers to the Adoption of Instructional Technology.

Belanger, A. L., Diekman, A. B., & Steinberg, M. (2017). Leveraging communal experiences in the curriculum: Increasing interest in pursuing engineering by changing stereotypic expectations. *Journal of Applied Social Psychology*, *47*(6), 305-319.

Beneroso, D., & Erans, M. (2020). Team-based learning: an ethnicity-focused study on the perceptions of teamwork abilities of engineering students. European Journal of Engineering Education, 1-12.

Bezrukova, K., Spell, C. S., Perry, J. L., & Jehn, K. A. (2016). A meta-analytical integration of over 40 years of research on diversity training evaluation. *Psychological Bulletin*, *142*(11), 1227.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery. *Journal of personality and social psychology*, *81*(5), 828.

Blascovich, J., & Mendes, W. B. (2000). Challenge and threat appraisals: The role of affective cues.

Bloxham*, S., & West, A. (2004). Understanding the rules of the game: marking peer assessment as a medium for developing students' conceptions of assessment. *Assessment & Evaluation in Higher Education*, *29*(6), 721-733.

Boysen, G. A., & Vogel, D. L. (2009). Bias in the classroom: Types, frequencies, and responses. *Teaching of Psychology*, *36*(1), 12-17.

Boysen, G. A., Vogel, D. L., Cope, M. A., & Hubbard, A. (2009). Incidents of bias in college classrooms: Instructor and student perceptions. *Journal of Diversity in Higher Education*, *2*(4), 219.

Brannon, T. N., Carter, E. R., Murdock-Perriera, L. A., & Higginbotham, G. D. (2018). From backlash to inclusion for all: Instituting diversity efforts to maximize benefits across group lines. *Social issues and policy review*, *12*(1), 57-90.

Brindley, C., & Scoffield, S. (1998). Peer assessment in undergraduate programmes. *Teaching in higher education*, *3*(1), 79-90.

Brown, S., & Knight, P. (1994). *Assessing learners in higher education*. Psychology Press.

Brown, T. N., Tanner-Smith, E. E., Lesane-Brown, C. L., & Ezell, M. E. (2007). Child, parent, and situational correlates of familial ethnic/race socialization. *Journal of Marriage and Family*, *69*(1), 14-25.

Bryan, R. E., Krych, A. J., Carmichael, S. W., Viggiano, T. R., & Pawlina, W. (2005). Assessing professionalism in early medical education: experience with peer evaluation and self-evaluation in the gross anatomy course. *Annals-Academy of Medicine Singapore*, *34*(8), 486.

Burke, S. E., Dovidio, J. F., Przedworski, J. M., Hardeman, R. R., Perry, S. P., Phelan, S. M., ... & Van Ryn, M. (2015). Do contact and empathy mitigate bias against gay and lesbian people among heterosexual medical students? A report from medical student CHANGES. *Academic medicine: journal of the Association of American Medical Colleges*, *90*(5), 645.

Campbell, K. S., Mothersbaugh, D. L., Brammer, C., & Taylor, T. (2001). Peer versus self assessment of oral business presentation performance. Business Communication Quarterly, 64(3).

Carr, P. B., Dweck, C. S., & Pauker, K. (2012). "Prejudiced" behavior without prejudice? Beliefs about the malleability of prejudice affect interracial interactions. *Journal of personality and social psychology*, *103*(3), 452.

Castaño-Muñoz, J., Sancho-Vinuesa, T., & Duart, J. M. (2013). Online interaction in higher education: Is there evidence of diminishing returns?. International Review of Research in Open and Distributed Learning, 14(5), 240-257

Catharsis Productions. (2017, November 8). What are microaggressions [Video]. YouTube. https://www.youtube.com/watch?v=ho_WW7M5E3A

Chang, E. H., Milkman, K. L., Gromet, D. M., Rebele, R. W., Massey, C., Duckworth, A. L., & Grant, A. M. (2019). The mixed effects of online diversity training. *Proceedings of the National Academy of Sciences*, *116*(16), 7778-7783.

Chen, M., & Bargh, J. A. (1997). Nonconscious behavioral confirmation processes: The self-fulfilling consequences of automatic stereotype activation. *Journal of Experimental Social Psychology*, *33*(5), 541-560.

Cheng, W., & Warren, M. (1997). Having second thoughts: Student perceptions before and after a peer assessment exercise. *Studies in Higher Education*, *22*(2), 233-239.

Connor, P., & Evers, E. R. (2020). The bias of individuals (in crowds): Why implicit bias is probably a noisily measured individual-level construct. *Perspectives on Psychological Science*, *15*(6), 1329-1345.

Cox, J. A., & Krumboltz, J. D. (1958). Racial bias in peer ratings of basic airmen. *Sociometry*, *21*(4), 292-299.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of consulting psychology*, *24*(4), 349.

Dancer, D., & Kamvounias, P. (2005). Student involvement in assessment: A project designed to assess class participation fairly and reliably. *Assessment & Evaluation in Higher Education*, *30*(4), 445-454.

Dancer, W. T., & Dancer, J. (1992). Peer rating in higher education. *Journal of Education for Business*, *67*(5), 306-309.

Das, M. (1998). Self and tutor evaluations in problem-based learning tutorials: is there a relationship?. *Medical education*, *32*(4), 411-418.

Dawson, P. (2017). Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education*, *42*(3), 347-360.

Deeley, S. J., & Bovill, C. (2017). Staff student partnership in assessment: enhancing assessment literacy through democratic practices. *Assessment & Evaluation in Higher Education*, *42*(3), 463-477.

Dejung, J. E., & Kaplan, H. (1962). Some differential effects of race of rater and ratee on early peer ratings of combat aptitude. *Journal of Applied Psychology*, *46*(5), 370.

Devine, P. G., & Monteith, M. J. (1993). The role of discrepancy-associated affect in prejudice reduction. In *Affect, cognition and stereotyping* (pp. 317-344). Academic Press.

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of experimental social psychology*, *48*(6), 1267-1278.

Devine, P. G., Forscher, P. S., Cox, W. T., Kaatz, A., Sheridan, J., & Carnes, M. (2017). A gender bias habit-breaking intervention led to increased hiring of female faculty in STEMM departments. *Journal of Experimental Social Psychology*, *73*, 211-215.

DiAngelo, R. (2018). *White fragility: Why it's so hard for white people to talk about racism*. Beacon Press.

Dovidio, & R. H. Fazio, (1992). "New technologies for the direct and indirect assessment of attitudes," In Questions about survey questions: Meaning, memory, attitudes, and social interaction, J. Tanur, Ed. New York: Russell Sage Foundation, 1992, pp. 204-237.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of personality and social psychology*, *82*(1), 62.

Duguid, M. M., & Thomas-Hunt, M. C. (2015). Condoning stereotyping? How awareness of stereotyping prevalence impacts expression of stereotypes. *Journal of Applied Psychology*, *100*(2), 343.

Eagly, A. H., & Carli, L. L. (1981). Sex of researchers and sex-typed communications as determinants of sex differences in influenceability: a meta-analysis of social influence studies. *Psychological Bulletin*, *90*(1), 1.

Evans, A. W., Leeson, R. M., & Petrie, A. (2007). Reliability of peer and self-assessment scores compared with trainers' scores following third molar surgery. Medical Education, 41(9), 866-872.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of educational research*, *70*(3), 287-322.

Falchikov, N., & Magin, D. (1997). Detecting gender bias in peer marking of students' group process work. *Assessment & Evaluation in Higher Education*, *22*(4), 385-396.

Fine, E., Wendt, A., & Carnes, M. (2014). Gendered expectations: are we unintentionally undermining our efforts to diversify STEM fields?. *XRDS: Crossroads, The ACM Magazine for Students*, *20*(4), 46-51.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, *117*(3), 522.

Forscher, P. S., Mitamura, C., Dix, E. L., Cox, W. T., & Devine, P. G. (2017). Breaking the prejudice habit: Mechanisms, timecourse, and longevity. *Journal of experimental social psychology*, *72*, 133-146.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences, 111(23), 8410-8415.

Fusion Comedy. (2016, October 5). How microaggressions are like mosquito bites [Video]. YouTube. https://www.youtube.com/watch/hDd3bzA7450

Gaillet, L. L. (1992). A Foreshadowing of Modern Theories and Practices of Collaborative Learning: The Work of Scottish Rhetorician George Jardine.

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of personality and social psychology*, *78*(4), 708.

Gatfield, T. (1999). Examining student satisfaction with group projects and peer assessment. *Assessment & Evaluation in Higher Education*, *24*(4), 365-377.

Golightly, A. (2021). Self-and peer assessment of preservice geography teachers' contribution in problem-based learning activities in geography education. International Research in Geographical and Environmental Education, 30(1), 75-90.

Gollwitzer, P. M. (1999). Implementation intentions: strong effects of simple plans. *American psychologist*, *54*(7), 493.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.

Gunter, R., & Stambach, A. (2005). DIFFERENCES IN MEN AND WOMEN SCIENTISTS'PERCEPTIONS OF WORKPLACE CLIMATE. *Journal of Women and Minorities in Science and Engineering*, *11*(1).

Hadim, H. A., & Esche, S. K. (2002, November). Enhancing the engineering curriculum through project-based learning. In *32nd Annual Frontiers in Education* (Vol. 2, pp. F3F-F3F). IEEE.

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self-and peer-assessment: The students' views. *Higher Education Research & Development*, *20*(1), 53-70.

Harland, T., Wald, N., & Randhawa, H. (2017). Student peer review: enhancing formative feedback with a rebuttal. *Assessment & Evaluation in Higher Education*, *42*(5), 801-811.

Harris, L. R., & Brown, G. T. (2013). Opportunities and obstacles to consider when using peer-and self-assessment to improve student learning: Case studies into teachers' implementation. *Teaching and Teacher Education*, *36*, 101-111.

Henderson, C., Beach, A. & Finkelstein, N. (2011). Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *Journal of Research in Science Teaching, 48*(8), 952-984. doi: 10.1002/tea.20439

Herek, G. M., & Capitanio, J. P. (1996). " Some of My Best Friends" Intergroup Contact, Concealable Stigma, and Heterosexuals' Attitudes Toward Gay Men and Lesbians. *Personality and Social Psychology Bulletin*, *22*(4), 412-424.

Holmes, Andrew Gary Darwin. “Researcher Positionality - A Consideration of Its Influence and Place in Qualitative Research - A New Researcher Guide.” Shanlax International Journal of Education, vol. 8, no. 4, 2020, pp. 1-10

Ikizer, E. G., & Blanton, H. (2016). Media coverage of “wise” interventions can reduce concern for the disadvantaged. *Journal of experimental psychology: applied*, *22*(2), 135.

Indiana University Center for Postsecondary Research (2021). The Carnegie Classification of Institutions of Higher Education, 2021 edition, Bloomington, IN: Author.

Iverson, S. V., & Seher, C. (2014). Using theatre to change attitudes toward lesbian, gay, and bisexual students. *Journal of LGBT Youth*, *11*(1), 40-61.

Jaime, A., Blanco, J. M., Domínguez, C., Sánchez, A., Heras, J., & Usandizaga, I. (2016). Spiral and project-based learning with peer assessment in a computer science project management course. *Journal of Science Education and Technology*, *25*(3), 439-449.

Jefferson, H., & Lewis, N., Jr. (2018, April 23). Starbucks won’t have any idea whether its diversity training works. Washington Post. Retrieved from https://www.washingtonpost.com/ news/posteverything/wp/2018/04/23/ starbucks-wont-have-any-ideawhether-its-diversity-training-works/

Jordan, C. H., Spencer, S. J., & Zanna, M. P. (2003, January). I love me... I love me not: Implicit self-esteem, explicit self-esteem, and defensiveness. In *Motivated social perception: The Ontario symposium* (Vol. 9, pp. 117-145). Lawrence Erlbaum Associates, Inc. Mahwah, NJ.

Kaiser, C. R., Major, B., Jurcevic, I., Dover, T. L., Brady, L. M., & Shapiro, J. R. (2013). Presumed fair: Ironic effects of organizational diversity structures. *Journal of personality and social psychology*, *104*(3), 504.

Karaca, E. (2009). An evaluation of teacher trainee's opinions of the peer assessment in terms of some variables. *World Applied Sciences Journal*, *6*(1), 123-128.

Karimi, A., & Manteufel, R. (2020, July). Most Recent Updates to ABET-EAC-Criteria 3, 4 and 5. In *2020 Gulf Southwest Section Conference*.

Kaufman, D. B., Felder, R. M., & Fuller, H. (1999, June). Peer ratings in cooperative learning teams. In *Proceedings of the 1999 Annual ASEE Meeting* (pp. 1-11).

Kaufman, D. B., Felder, R. M., & Fuller, H. (2000). Accounting for individual effort in cooperative learning teams. *Journal of Engineering Education*, *89*(2), 133-140.

Kelley, L., Chou, C. L., Dibble, S. L., & Robertson, P. A. (2008). A critical intervention in lesbian, gay, bisexual, and transgender health: knowledge and attitude outcomes among second-year medical students. *Teaching and learning in medicine*, *20*(3), 248-253.

Kennell, Vicki; Elliot, Amy; and Weirick, Joshua, "Tinkering with Comments: Tailoring Practice by Spying on Written Artifacts" (2017). Purdue Writing Lab/Purdue OWL Presentations. Paper 15. https://docs.lib.purdue.edu/writinglabpres/15

Kim, H. S. (2014). Uncertainty analysis for peer assessment: oral presentation skills for final year project. *European Journal of Engineering Education*, *39*(1), 68-83.

Knobloch, N. A., & Ball, A. L. (2006, April). Analyzing the contextual, motivational, and conceptual characteristics of teaching faculty in regard to the use of learner centered approaches to teaching. In *annual meeting of the American Education Research Association, San Francisco, CA*.

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of applied psychology*, *70*(1), 56.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological bulletin*, *87*(1), 72.

Langan, A. M., Wheater, C. P., Shaw, E. M., Haines, B. J., Cullen, W. R., Boyle, J. C., ... & Preziosi, R. F. (2005). Peer assessment of oral presentations: effects of student gender, university affiliation and participation in the development of assessment criteria. *Assessment & Evaluation in Higher Education*, *30*(1), 21-34.

Lee, K., Quinn, P. C., & Heyman, G. D. (2017). Rethinking the emergence and development of implicit racial bias: A perceptual-social linkage hypothesis. *New perspectives on human development*, 27-46.

Li, L., & Gao, F. (2016). The effect of peer assessment on project performance of students at different learning levels. *Assessment & Evaluation in Higher Education*, *41*(6), 885-900.

Lima, Pernille Hammar Andersson & Elisabeth Saalman (2017) Active Learning in Engineering Education: a (re)introduction, European Journal of Engineering Education, 42:1, 1-4, DOI: 10.1080/03043797.2016.1254161

Lind, D. S., Rekkas, S., Bui, V., Lam, T., Beierle, E., & Copeland Iii, E. M. (2002). Competency-based student self-assessment on a surgery rotation. *Journal of Surgical Research*, *105*(1), 31-34.

Liow, J. L. (2008). Peer assessment in thesis oral presentation. *European Journal of Engineering Education*, *33*(5-6), 525-537.

López-Pastor, V., & Sicilia-Camacho, A. (2017). Formative and shared assessment in higher education. Lessons learned and challenges for the future. *Assessment & Evaluation in Higher Education*, *42*(1), 77-97.

Loughry, M. L., Ohland, M. W., & Woehr, D. J. (2014). Assessing teamwork skills for assurance of learning using CATME team tools. *Journal of Marketing Education*, *36*(1), 5-19.

Maass, A., Castelli, L., & Arcuri, L. (2000). Measuring prejudice: Implicit versus explicit techniques. *Social identity processes: Trends in theory and research*, 96-116.

Magin, D. (2001). Reciprocity as a source of bias in multiple peer assessment of group work. *Studies in Higher Education*, *26*(1), 53-63.

Magin, D., & Helmore, P. (2001). Peer and teacher assessments of oral presentation skills: how reliable are they?. *Studies in Higher Education*, *26*(3), 287-298.

Marcus, A., Mullins, L. C., Brackett, K. P., Tang, Z., Allen, A. M., & Pruett, D. W. (2003). Perceptions of racism on campus. *College Student Journal*, *37*(4), 611-627.

Marzano, R. J. (2010). Designing & teaching learning goals & objectives. Solution Tree Press.

Mathews, B. P. (1994). Assessing individual contributions: Experience of peer evaluation in major group projects. *British Journal of Educational Technology*, *25*(1), 19-28.

May, G. L. (2008). The effect of rater training on reducing social style bias in peer evaluation. *Business Communication Quarterly*, *71*(3), 297-313.

May, G. L., & Gueldenzoph, L. E. (2006). The effect of social style on peer evaluation ratings in project teams. *The Journal of Business Communication (1973)*, *43*(1), 4-20.

McConahay, (1986). Modern racism, ambivalence, and the Modern Racism Scale, in Prejudice, discrimination, and racism, J. F. Dovidio & S. L. Gaertner, Eds. San Diego, CA: Academic Press, 1986, pp. 91-125.

McInerney, M. J., & Fink, L. D. (2003). Team-based learning enhances long-term retention and critical thinking in an undergraduate microbial physiology course. Microbiology education, 4(1), 3-12.

McLeod, B. D., Jensen-Doss, A., & Ollendick, T. H. (Eds.). (2013). *Diagnostic and behavioral assessment in children and adolescents: A clinical guide*. Guilford Press.

Michaelsen, L. K., Knight, A. B., & Fink, L. D. (2004). Team-based learning: A transformative use of small groups in college teaching.

Michaelsen, L. K., Sweet, M., & Parmelee, D. X. (Eds.). (2011). *Team-Based Learning: Small Group Learning's Next Big Step: New Directions for Teaching and Learning, Number 116* (Vol. 103). John Wiley & Sons.

Miedijensky, S., & Tal, T. (2009). Embedded Assessment in Project-based Science Courses for the Gifted: Insights to inform teaching all students. *International Journal of Science Education*, *31*(18), 2411-2435.

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. *Journal of Personality and Social Psychology*, *65*(3), 469.

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition*, *19*(4), 395-417.

Montgomery, B. M. (1986). An interactionist analysis of small group peer assessment. *Small Group Behavior*, *17*(1), 19-37.

Moorman, D., & Wicks-Smith, D. (2012). Poverty discrimination revealed through student peer evaluations. *College Student Journal*, *46*(1), 141-149.

Morrison, M. A., & Morrison, T. G. (2003). Development and validation of a scale measuring modern prejudice toward gay men and lesbian women. *Journal of homosexuality*, *43*(2), 15-37.

Najdanovic-Visak, V. (2017). Team-based learning for first year engineering students. *Education for Chemical Engineers*, *18*, 26-34.

National Academy of Engineering. (2015). NAE Grand Challenges for Engineering. Retrieved from http://engineeringchallenges.org.

National Research Council (NRC). (2012). *Discipline-based educational research: Understanding and improving learning in undergraduate science and engineering*. Washington, DC: National Academies Press

Nevid, J. S., & McClelland, N. (2010). Measurement of implicit and explicit attitudes toward Barack Obama. *Psychology & Marketing*, *27*(10), 989-1000.

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education*, *39*(1), 102-122.

Ohland, M. W., Loughry, M. L., Woehr, D. J., Bullard, L. G., Felder, R. M., Finelli, C. J., ... & Schmucker, D. G. (2012). The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self-and peer evaluation. *Academy of Management Learning & Education*, *11*(4), 609-630.

O'Neill, G. P. (1985). Self, teacher and faculty assessments of student teaching performance: a second scenario. *Alberta journal of educational research*, *31*(2), 88-98.

ONeill, T. A., Boyce, M., & McLarnon, M. J. (2020, May). Team health and project quality are improved when peer evaluation scores affect grades on team projects. In Frontiers in Education (Vol. 5, p. 49). Frontiers Media SA.

Pajares, F., & Johnson, M. J. (1996). Self-efficacy beliefs and the writing performance of entering high school students. *Psychology in the Schools*, *33*(2), 163-175.

Pantos, A. J., & Perkins, A. W. (2013). Measuring implicit and explicit attitudes toward foreign accented speech. *Journal of Language and Social Psychology*, *32*(1), 3-20.

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of social distance? Failure to replicate effects on social and food judgments.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*(4), 233-248.

Planas Lladó, A., Soley, L. F., Fraguell Sansbelló, R. M., Pujolras, G. A., Planella, J. P., Roura-Pascual, N., ... & Moreno, L. M. (2014). Student perceptions of peer assessment: an interdisciplinary study. *Assessment & Evaluation in Higher Education*, *39*(5), 592-610.

Plant, E. A., Devine, P. G., Cox, W. T., Columb, C., Miller, S. L., Goplen, J., & Peruche, B. M. (2009). The Obama effect: Decreasing implicit prejudice and stereotyping. *Journal of Experimental Social Psychology*, *45*(4), 961-964.

Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. *Journal of Applied Psychology*, *74*(5), 770.

Rafiq, Y., & Fullerton, H. (1996). Peer assessment of group projects in civil engineering. *Assessment and Evaluation in Higher Education*, *21*(1), 69-81.

Rees, C. (2003). Self-assessment scores and gender. *Medical Education*, *37*(6), 572-573.

Riess, M., Kalle, R. J., & Tedeschi, J. T. (1981). Bogus pipeline attitude assessment, impression management, and misattribution in induced compliance settings. *The Journal of Social Psychology*, *115*(2), 247-258.

Sackett, P. R., & DuBois, C. L. (1991). Rater-ratee race effects on performance evaluation: Challenging meta-analytic conclusions. *Journal of Applied Psychology*, *76*(6), 873.

Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning. *Educational assessment*, *11*(1), 1-31.

Samuel, E. (2004). Racism in peer-group interactions: South Asian students' experiences in Canadian academe. *Journal of College Student Development*, *45*(4), 407-424.

Schmidt, F. L., & Johnson, R. H. (1973). Effect of race on peer ratings in an industrial situation. *Journal of Applied Psychology*, *57*(3), 237.

Schultz, J. R., Gaither, S. E., Urry, H. L., & Maddox, K. B. (2015). Reframing anxiety to encourage interracial interactions. *Translational Issues in Psychological Science*, *1*(4), 392.

Schwan, S., & Riempp, R. (2004). The cognitive benefits of interactive videos: Learning to tie nautical knots. Learning and instruction, 14(3), 293-305.

Scruggs, T. E., Mastropieri, M. A., & Boon, R. (1998). Science education for students with disabilities: A review of recent research.

Sherrard, W. R., Raafat, F., & Weaver, R. R. (1994). An empirical study of peer bias in evaluations: Students rating students. *Journal of Education for Business*, *70*(1), 43-47.

Slavin, R. E., & Oickle, E. (1981). Effects of cooperative learning teams on student achievement and race relations: Treatment by race interactions. *Sociology of Education*, 174-180.

Slemmons, K., Anyanwu, K., Hames, J., Grabski, D., Mlsna, J., Simkins, E., & Cook, P. (2018). The impact of video length on learning in a middle-level flipped science setting: Implications for diversity inclusion. Journal of Science Education and Technology, 27(5), 469-479

Smith, H., Cooper, A., & Lancaster, L. (2002). Improving the quality of undergraduate peer assessment: A case for student and staff development. *Innovations in education and teaching international*, *39*(1), 71-81.

Snowball, J. D., & Mostert, M. (2013). Dancing with the devil: Formative peer assessment and academic performance. *Higher Education Research & Development*, *32*(4), 646-659.

Snowden, J. L. (2005). *Explicit and implicit bias measures: Their relation and utility as predictors of criminal verdict tendency* (Doctoral dissertation, University of North Carolina at Wilmington).

Spence, J. T., Helmreich, R., & Stapp, J. (1973). A short version of the Attitudes toward Women Scale (AWS). *Bulletin of the Psychonomic society*, *2*(4), 219-220.

Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of educational research*, *69*(1), 21-51.

Staats, K. Capatosto, R. A. Wright, & D. Contractor. (2014). State of the science: Implicit bias review 2014. *Kirwan Institute for the Study of Race and Ethnicity*, vol. 14, pp. 129-142

Stefani, L. A. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education*, *19*(1), 69-75.

Strijbos, J. W., Ochoa, T. A., Sluijsmans, D. M., Segers, M. S., & Tillema, H. H. (2009). Fostering interactivity through formative peer assessment in (web-based) collaborative learning environments. In *Cognitive and emotional processes in web-based education: Integrating human factors and personalization* (pp. 375-395). IGI Global.

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., ... & Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. Proceedings of the National Academy of Sciences, 117(12), 6476-6483.

Thondhlana, G., & Belluigi, D. Z. (2017). Students' reception of peer assessment of group-work contributions: Problematics in terms of race and gender emerging from a South African case study. *Assessment & Evaluation in Higher Education*, *42*(7), 1118-1131.

Todd, A. R., Bodenhausen, G. V., Richeson, J. A., & Galinsky, A. D. (2011). Perspective taking combats automatic expressions of racial bias. *Journal of personality and social psychology*, *100*(6), 1027.

Tolmie, A. K., Topping, K. J., Christie, D., Donaldson, C., Howe, C., Jessiman, E., ... & Thurston, A. (2010). Social effects of collaborative learning in primary schools. *Learning and instruction*, *20*(3), 177-191.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of educational Research*, *68*(3), 249-276.

Topping, K. J. (2010). Methodological quandaries in studying process and outcomes in peer assessment. *Learning and Instruction*, *20*(4), 339-343.

Topping, K. J. (2013). Peers as a source of formative and summative assessment. In *SAGE handbook of research on classroom assessment* (pp. 395-412). Sage Publications.

Trawalter, S., & Richeson, J. A. (2008). Let's talk about race, baby! When Whites' and Blacks' interracial contact experiences diverge. *Journal of Experimental Social Psychology*, *44*(4), 1214-1217.

Trevelyan, J. (2014). *The making of an expert engineer*. Crc Press.

Tucker, R. (2014). Sex does not matter: gender bias and gender differences in peer assessments of contributions to group work. *Assessment & Evaluation in Higher Education*, *39*(3), 293-309.

Twenge, J. M. (1997). Attitudes toward women, 1970–1995: A meta-analysis. *Psychology of Women Quarterly*, *21*(1), 35-51.

València, H. O., Carrillo, Á. G., & Benítez, M. G. (2012). The influence of social style in evaluating academic presentations of engineering projects. *Journal of Technology and Science Education*, *2*(2), 68-76.

Van den Berg, I., Admiraal, W., & Pilot, A. (2006). Designing student peer assessment in higher education: Analysis of written and oral peer feedback. *Teaching in higher education*, *11*(2), 135-147.

Van Zundert, M., Sluijsmans, D., & Van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and instruction*, *20*(4), 270-279.

Van-Trieste, R. F. (1990). The relation between Puerto Rican university students' attitudes toward Americans and the students' achievement in English as a Second Language. *Homines 13*, *14*, 94-112.

Wald, J. (2014). Can "de-biasing" strategies help to reduce racial disparities in school discipline. *A summary of the literature," Discipline Disparities: A Research-to-Practice Collaborative*.

Wang, J., & Imbrie, P. K. (2010). Students' peer evaluation calibration through the administration of vignettes. In *American Society for Engineering Education*. American Society for Engineering Education.

Wayland, C., Walker, L., & Ferrara, A. (2014). Opening the Black Box: How Gender and Race Mediate Collaborative Learning in the Classroom. In *the Southeastern Women's Studies Association Annual Meeting, Wilmington, NC, USA*.

Wen, M. L., & Tsai, C. C. (2006). University students' perceptions of and attitudes toward (online) peer assessment. *Higher Education*, *51*(1), 27-44.

Widjaja, W., & Takahashi, M. (2016). Distributed interface for group affinity-diagram brainstorming. Concurrent Engineering, 24(4), 344-358.

Wilder, D. A. (1978). Reduction of intergroup discrimination through individuation of the out-group. *Journal of Personality and Social Psychology*, *36*(12), 1361.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological review*, *107*(1), 101.

Witte, K. (1992). Putting the fear back into fear appeals: The extended parallel process model. *Communications Monographs*, *59*(4), 329-349.

Xiao, W. S., Fu, G., Quinn, P. C., Qin, J., Tanaka, J. W., Pascalis, O., & Lee, K. (2015). Individuation training with other-race faces reduces preschoolers' implicit racial bias: A link between perceptual and social representation of faces in children. *Developmental Science, 18*(4), 655-663.

Xu, K., Nosek, B., & Greenwald, A. (2014). Psychology data from the race implicit association test on the project implicit demo website. *Journal of Open Psychology Data*, *2*(1).

APPENDIX A: FINAL TRAINING INTRODUCTION VIDEO SCRIPT

You might be wondering why you're here. Peer assessment seems pretty straightforward, right? You just give your peers a score and click submit. Done!

However, not everyone experiences peer assessment and teamwork the same way. In a survey of over 300 students, 18 percent had or might have experienced unfairness in their peer assessments, and 46 percent had or might have experienced unfairness from their teammates. Working in teams is one of the most effective ways to learn, so it's important that everyone has a fair experience.

This training will show you how to assess your peers fairly. You'll learn how to write constructive, helpful feedback while avoiding being unfair to your classmates. Of course, your classmates will learn the same thing! That means that you'll be giving AND getting better peer assessments. Everyone (including you and your grade in the class!) benefits here.

Along the way, you'll also learn how to be a better communicator. As we all know, communication skills are essential, no matter your major. Knowing how to communicate and give feedback to your peers and coworkers can set you apart from other job candidates, and could even help you land that dream job.

You'll be here for about 30 minutes. You can pause the training at any time and start where you left off, just make sure you're using the same device you started on. Most pages of the training are timed – the "next" button won't appear until you've had enough time to watch the video or read the content. You'll start off learning about how to make feedback helpful and then learn how to make sure your helpful feedback is fair. As you go, you'll watch a few more videos, answer questions, and write example peer assessment comments. When you're done, you should be an expert peer assessor!

APPENDIX B: FINAL TRAINING HELPFUL FEEDBACK VIDEO SCRIPT

How do you write feedback that your peers can use? In this video, we'll talk about seven things you should consider when writing feedback. These seven things aren't just useful for peer assessments, they're useful for any kind of feedback you give! Remember, feedback can be both positive AND constructive. It's important to make sure both types of feedback are useful.

Feedback should be about the process, not the person. Feedback isn't about passing judgement, like calling someone rude or saying that a presentation was boring. For example, imagine your classmate gave a presentation that you thought was too long. Instead of saying "you made your presentation too long; it bored me", you could say "Instead of 3 examples per slide, you could limit to 1 or 2 examples per slide. That would make the presentation more succinct and impactful, and reduce the length!"

Feedback should be specific – your reader needs to know what you're talking about! Imagine your teammate set up all the folders your team needed for a project on Google Drive. Everyone thought the folder system was super helpful and kept them organized. You could say "good job on the project", but it would be a lot more useful to be specific and say "the folder system you made helped the team work more efficiently, well done!"

Feedback should be honest and sincere – it should sound like YOU wrote it and get right to the point. Your classmate Destiny is very enthusiastic, but interrupts A LOT. It's gotten to the point that some of your teammates hesitate to speak at all for fear of interruption. You consider just ignoring it, but that wouldn't help anyone. Instead, you give her honest feedback in her peer evaluation: "Interrupting makes others hesitant to contribute. Your enthusiasm is great, but in the future try not to cut anyone off." Especially when giving constructive feedback, following a "cause" (interrupting), "effect", (makes others hesitate to speak), "future action" (try not to cut anyone off) format can reduce feelings of defensiveness by the receiver.

Feedback should be actionable. When someone reads your feedback, they should be able to answer the question: What specifically should I do more of and less of next time, based on this information? We often think of actionable feedback being about things we would like someone to change, but it's important to tell others what they're doing well and should continue. For example "good job" is pleasant, but isn't actionable. On the other hand "the folder system you made helped us work more efficiently, it would be great if you could do that for the next project too!" is both positive AND actionable.

Feedback should be timely! The goal of feedback is for the receiver to reflect upon and change or consciously continue a behavior. If feedback is given too late, the window for reflecting on that behavior is already closed. Prioritize giving feedback in the peer assessment closest to when the behavior happened – if it happened at the beginning of the semester, don't wait until the end of the semester to talk about it!

When giving feedback, it's important that the receiver is in a position to receive it. Imagine you're hurrying to another class to take an exam when your teammate offers you comments about a presentation. It's likely that the feedback won't really "sink in" or could even make you annoyed with your teammate. A better solution is to give feedback during the peer assessment period or to ask the receiver if they're in a position to hear your comments. This way, the recipient is prepared and ready for what you have to say.

One of the most important things to remember is that feedback should be concerned with behavior that the receiver can change. It isn't helpful to give feedback on someone's accent, age,

race, or previous experience. Even if the receiver wanted to change these attributes, they couldn't. Additionally, it isn't helpful to give feedback on how you assume someone should act. For example, imagine one of your teammates is from Brazil. Commenting "Your English is really good!" on a peer assessment indicates that you had assumed their English wouldn't be good.

To help you out with things you should consider in your evaluations, here are some ideas. Participation: attendance at class or meetings, contributing ideas to a discussion, usefulness of their contributions, attention to class or meetings. Preparation: Like completion of pre-work and equipped with the supplies they need. Reliability: have they completed the tasks assigned to them in a timely manner? Have they communicated about any changes to their schedule or workflow? And finally, Teamwork: listening to the views of others, their attitude toward teamwork, and their level of cooperation.

These tips should give you the foundation you need to give fair and useful feedback!

APPENDIX C: IRB 19-295

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Research
Vice President for Research
2420 Lincoln Way, Suite 202
Ames, Iowa 50014
515 294-4566

Date: 06/19/2019

To: Jacklin Stonewall Michael Dorneich, Ph.D.

From: Office for Responsible Research

Title: **Student Attitudes Toward Peer Assessment**

IRB ID: **19-295**

Submission Type: Initial Submission **Exemption Date:** 06/19/2019

The project referenced above has been declared exempt from most requirements of the human subject protections regulations as described in 45 CFR 46.104 or 21 CFR 56.104 because it meets the following federal requirements for exemption:

2018 - 2 (iii): Research that only includes interactions involving educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior (including visual or auditory recording) when the information obtained is recorded by the investigator in such a manner that the identity of the human subjects can readily be ascertained, directly or through identifiers linked to the subjects, and an IRB conducts a LIMITED IRB REVIEW to [determine there are adequate provisions to protect the privacy of subjects and to maintain confidentiality of the data].

The determination of exemption means that:

- **You do not need to submit an application for continuing review. Instead, you will receive a request for a brief status update every three years. The status update is intended to verify that the study is still ongoing.**
- **You must carry out the research as described in the IRB application.** Review by IRB staff is required prior to implementing modifications that may change the exempt status of the research. In general, review is required for any *modifications to the research procedures* (e.g., method of data collection, nature or scope of information to be collected, nature or duration of behavioral interventions, use of deception, etc.), any change in *privacy or confidentiality protections*, modifications that result in the *inclusion of participants from vulnerable populations*, removing plans for informing participants about the study, any *change that may increase the risk or discomfort to participants, and/or* any change such that the revised procedures do not fall into one or more of the regulatory exemption categories. The purpose of review is to determine if the project still meets the federal criteria for exemption.
- All ***changes to key personnel*** must receive prior approval.
- **Promptly inform the IRB of any addition of or change in federal funding for this study.** Approval of the protocol referenced above applies only to funding sources that are specifically identified in the corresponding IRB application.

APPENDIX D: IRB 19-516

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Research
Vice President for Research
2420 Lincoln Way, Suite 202
Ames, Iowa 50014
515 294-4566

Date: 10/17/2019

To: Jacklin Stonewall Michael Dorneich, Ph.D.

From: Office for Responsible Research

Title: **Instructor Attitudes toward Peer Assessment**

IRB ID: **19-516**

Submission Type: Initial Submission **Exemption Date:** 10/17/2019

The project referenced above has been declared exempt from most requirements of the human subject protections regulations as described in 45 CFR 46.104 or 21 CFR 56.104 because it meets the following federal requirements for exemption:

2018 - 2 (ii): Research that only includes interactions involving educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior (including visual or auditory recording) when any disclosure of the human subjects' responses outside the research would not reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, educational advancement, or reputation.

The determination of exemption means that:

- **You do not need to submit an application for continuing review. Instead, you will receive a request for a brief status update every three years. The status update is intended to verify that the study is still ongoing.**
- **You must carry out the research as described in the IRB application.** Review by IRB staff is required prior to implementing modifications that may change the exempt status of the research. In general, review is required for any *modifications to the research procedures* (e.g., method of data collection, nature or scope of information to be collected, nature or duration of behavioral interventions, use of deception, etc.), any change in *privacy or confidentiality protections*, modifications that result in the *inclusion of participants from vulnerable populations*, removing plans for informing participants about the study, any *change that may increase the risk or discomfort to participants, and/or* any change such that the revised procedures do not fall into one or more of the regulatory exemption categories. The purpose of review is to determine if the project still meets the federal criteria for exemption.
- All ***changes to key personnel*** must receive prior approval.
- **Promptly inform the IRB of any addition of or change in federal funding for this study.** Approval of the protocol referenced above applies only to funding sources that are specifically identified in the corresponding IRB application.

APPENDIX E: IRB 20-055

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Research
Vice President for Research
2420 Lincoln Way, Suite 202
Ames, Iowa 50014
515 294-4566

Date: 02/28/2020

To: Michael Dorneich, Ph.D.

From: Office for Responsible Research

Title: Peer Assessment Training - Bias Reduction

IRB ID: 20-055

Submission Type: Initial Submission **Exemption Date:** 02/28/2020

The project referenced above has been declared exempt from most requirements of the human subject protections regulations as described in 45 CFR 46.104 or 21 CFR 56.104 because it meets the following federal requirements for exemption:

2018 - 1: Research, conducted in established or commonly accepted educational settings, that specifically involves normal educational practices that are not likely to adversely impact students' opportunity to learn required educational content or the assessment of educators who provide instruction. This includes most research on regular and special education instructional strategies, and research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods.
2018 - 2 (ii): Research that only includes interactions involving educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior (including visual or auditory recording) when any disclosure of the human subjects' responses outside the research would not reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, educational advancement, or reputation.

The determination of exemption means that:

- **You do not need to submit an application for continuing review. Instead, you will receive a request for a brief status update every three years. The status update is intended to verify that the study is still ongoing.**
- **You must carry out the research as described in the IRB application.** Review by IRB staff is required prior to implementing modifications that may change the exempt status of the research. In general, review is required for any *modifications to the research procedures* (e.g., method of data collection, nature or scope of information to be collected, nature or duration of behavioral interventions, use of deception, etc.), any change in *privacy or confidentiality protections*, modifications that result in the *inclusion of participants from vulnerable populations*, removing plans for informing participants about the study, any *change that may increase the risk or discomfort to participants, and/or* any change such that the revised procedures do not fall into one or more of the regulatory exemption categories. The purpose of review is to determine if the project still meets the federal criteria for exemption.