Higher Education (2022) 83:1315–1333
https://doi.org/10.1007/s10734-021-00744-1

Gender-biased evaluation or actual differences? Fairness in the evaluation of faculty teaching

Edgar Valencia[1]

Accepted: 21 July 2021 / Published online: 13 August 2021

Abstract

How do we know if a faculty teaching evaluation is biased? *Biasing factors* studies are an influential source of evidence for arguing about biased teaching evaluations. These studies examine existing evaluation data and compare the results by gender, race, or ethnicity, interpreting a significant difference between subgroups as evidence of bias. However, only a difference explained by irrelevant aspects embedded in the evaluation would compromise its fairness. The study aims to amend how practitioners and researchers address gender bias concerns in faculty teaching evaluations by defining *fairness*, *disparate impact*, and *statistical bias* from an educational measurement standpoint. The study illustrates the use of differential item functioning (DIF) analysis, a strategy to examine whether the meaning of an item changes depending on the gender of the instructor. The study examines instructor's gender bias using responses to a course evaluation questionnaire from education graduate students from two academic departments within the same institution. In one of the departments, the analysis suggested a fair evaluation and no gender gap. In the other department, four of the eight items in the rating scale were easier for women than men with similar teaching ability, and women achieved better evaluations than men. The discussion addresses the conceptual and methodological advantages of adopting an educational measurement perspective on fairness in faculty teaching evaluation. Findings encourage practitioners and administrators to use the best available tools to strengthen the credibility of faculty teaching evaluations and prevent unfair personnel decisions affecting underrepresented subgroups in academia by gender, race, or ethnicity.

Keywords Bias · Fairness · Gender gap · Faculty teaching evaluation · Validity · Disparate impact

✉ Edgar Valencia
envalenc@uc.cl

[1] Facultad de Educación, Pontificia Universidad Católica de Chile, Avenida Vicuña Mackenna 4860, Macul, Santiago, Chile

The evaluation of faculty teaching faces constant scrutiny of the various aspects that affect validity. Research on student evaluation of teaching (SET) calls into question the number and nature of the teaching attributes included in these questionnaires (Alhija, 2017; Spooren et al., 2013), the quality in which students engage the response process (Bassett et al., 2017; McClain et al., 2017; Valencia, 2020) and the lack of correlation with students' grades (Uttl et al., 2017). Teaching evaluation correlates with variables unrelated to teaching ability, including personal characteristics such as gender (Andersen & Miller, 1997; Basow & Martin, 2012; Feldman, 1993; Stark & Freishtat, 2014; Boring et al., 2016). These findings feed a growing concern about gender-biased teaching evaluation (ASA, 2019; Gómez Cama et al., 2016; Mitchel & Martin, 2018; Weisshaar, 2017). The effects of gender bias in faculty teaching evaluation increase as the COVID-19 pandemic forces adaptations in working conditions, including online lectures, new grading systems, and fewer available resources, all more likely to affect women (Malisch et al., 2020)

The concept "glass ceiling effect" captures women's lack of access to better wages, power, and opportunities compared to men (Bertrand, 2017). The steps taken to prevent or minimize gender inequality and the glass ceiling effect in academia have produced slow results (Cundiff et al., 2018; Gómez Cama et al., 2016). One explanation of gender inequality in academia proposes that women invest less in education, training, and work experiences and accumulate less human capital than men because women seek a balance between work expectations and family obligations. Less human capital impacts productivity and puts women at a disadvantage when applying for academic jobs, tenure, leadership, and senior management positions. As a result, women are more often recruited by less elite institutions in less prestigious disciplines than men, they are less likely to receive tenure, and they progress slowly than men in their careers with lower wages (Gómez Cama et al., 2016; Weisshaar, 2017).

A second reason for gender inequality in academia relates to gender-biased evaluation (Weisshaar, 2017). Two mechanisms examined in faculty teaching evaluation literature are stereotyping (Arbuckle & Williams, 2003; Cundiff et al., 2018, Martin & Mitchell, 2018; Bavishi et al., 2010; Rivera & Tilcsik, 2019) and expectations violation (Anderson & Smith, 2005; Basow & Montgomery, 2005; MacNell et al., 2015).

A stereotype is a belief that shapes the judgments individuals make about members of a specific group based merely on group membership. In academia, stereotypes for women include warmth, nurturance, emotional sensitivity, and other similar terms. Stereotypes for men include competence, dominance, high status, authority, brilliance, and alike. Students may automatically use a gender stereotype to judge their instructors when filling a teaching evaluation form (Arbuckle & Williams, 2003). Stereotyping would result in higher scores for the group that better resemble the criteria utilized in the teaching evaluation, which often relates to "competence," hence favoring men. The evaluation itself may trigger gender stereotypes of competence, providing an advantage to men over women (Rivera & Tilcsik, 2019).

The second form of gender-biased teaching evaluation involves a contrast between behaviors and expectations. An expectation violation occurs when students hold beliefs not coherent with how their teacher behaves (MacNell et al., 2015). For instance, students may believe that availability is prototypical of a specific gender (women), profession (e.g., nursing, social work, education), or occupation (e.g., teacher). When a female teacher shows less availability to meet after class than expected by students, students may rate this teacher worse than a male teacher showing the same availability because she violated the expectations hold by her students.

Women may struggle to comply with conflicting gender and occupational expectations leading to a double bind threat (Cundiff et al., 2018). First, women may try to fulfill occupational expectations that allow progress in their careers. Examples of these expectations are assertiveness, self-promotion, and research roles. Women doing so risk punishment because these are expectations related to men. Second, women may avoid fulfilling attributes targeting service (e.g., supportiveness, nurturing, and teaching roles) that are less relevant for career progression. Again, women may suffer punishment because they are not meeting occupation expectations. Thus, stereotypes and expectations may shape students' judgment of their teachers and teaching, affecting responses to a teaching evaluation form.

How do we currently know when a teaching evaluation is gender-biased?

An assumption underlying the current discussion about gender bias in teaching evaluation concerns the validity and relevance of the research findings for administrators and practitioners. An influential first group of studies reports findings from experimental designs. Significant challenges relate to manipulating gender (or other characteristics such as age, race, or ethnicity) and the randomization of students to different study conditions (e.g., sections). For instance, Arbuckle and Williams (2003) asked college students to watch a 35-min picture-slide-audiotaped presentation where the instructor's figure and voice were neutral. Anderson and Smith (2005) changed the name (either feminine or masculine) and ethnicity of an instructor's CV before asking undergraduate students to rate his/her capability. Similarly, Bavishi et al. (2010) also manipulated the instructor's CV to create conditions based on gender, ethnicity, and discipline. Then, they asked college students to rate the instructor's competence. MacNell et al. (2015) conducted a similar manipulation by changing the instructor's name in an online introductory-level anthropology/sociology course. Finally, Bonitz (2011) manipulated gender using a hypothetical instructor and teaching situation described in a vignette. Due to manipulation, instructors differ in only one attribute (gender) and are comparable in all other attributes, including teaching ability. Students across conditions are comparable due to randomization. Holding constant teaching ability and students' severity is the basis for adequately inferring that a gender bias in teaching evaluation occurred. However, the artificial manipulation of gender reduces the generalizability of the findings over actual teaching conditions. Additionally, experiments are not helpful for administrators and practitioners seeking to examine gender bias because manipulating gender or randomly assigning students to course sections is unpractical.

Biasing factors literature

A second influential group of studies exploring gender bias in teaching evaluation relate to the *biasing factors* literature. Most SET research syntheses include a section on *biasing factors* (Alhija, 2017; Marsh, 1987; Onwuegbuzie et al., 2009; Spooren et al., 2013; Stark & Freishtat, 2014; Wachtel, 1998). An advantage of a *biasing factors* study over an experiment is that the former utilizes routinely collected teaching evaluation data (MacNell et al., 2015). This vein of literature defines a *bias* to occur when "a student, teacher, or course characteristic affects the evaluation made, either positively or negatively, but is unrelated to any criteria of good teaching" (Centra, 2003, p. 498)." Typical methods for determining a positive or negative *impact* are correlation analysis, regression analysis, and ANOVA. Any statistically significant finding (correlation coefficient, regression coefficient, or ANOVA main or interaction effect) would indicate a *biased* evaluation.

From the numerous examples of *biasing factors* studies addressing instructor's characteristics as a source of bias, findings are inconclusive or contradictory (Basow & Martin, 2012; Centra & Gaubatz, 2000; Spooren et al., 2013). Similarly, studies comparing women and men in teaching evaluation show inconsistent results, with women achieving higher evaluations than men on occasions. For instance, Basow and Montgomery (2005) report a statistically significant main effect of gender, with women receiving higher scores from students at a liberal arts college. Smith et al. (2007) report a statistically significant regression coefficient of gender on teaching evaluation scores, with undergraduate communication students scoring men higher than women. McPherson and colleagues report a statistically significant regression coefficient of being a male instructor on evaluation scores using responses from students attending undergraduate economy courses (McPherson et al., 2009). However, there is no effect when analyzing evaluations from economy master's students (McPherson & Jewell, 2007). Despite the prior belief that the gender effect is either null or small or that findings are inconclusive (Aleamoni & Hexner, 1980; Andersen & Miller, 1997; Centra & Gaubatz, 2000; Feldman 1993; Marsh & Roche, 1997; Ory, 2001; Theall & Franklin, 2001, Wachtel, 1998), recent evidence tends to support the concern about gender-biased teaching evaluations (ASA, 2018; Gómez Cama et al., 2016; Mengel, Sauermann, & Zölitz, 2019; Mitchel & Martin 2018; Rivera & Tilcsik, 2019; Wagner, Rieger, & Voorvelt, 2016, Weisshaar, 2017).

A biasing factors study does not measure bias

One general problem affecting *biasing factors* studies is that the relationship captured by a correlation or regression coefficient is not evidence of bias (Centra, 2003; Haladyna & Hess, 1994; MacNell et al., 2015; Marsh, 1987). These studies can indicate the direction and strength of the *effect* (or impact) of an independent variable (i.e., instructor's gender, race, or ethnicity) on a dependent variable (teaching evaluation) only if the study meets the conditions for inferring causality (Schneider et al., 2007; Shadish et al., 2002). Experimentation is the more effective strategy to achieve these conditions. Some studies introduce statistical controls (covariates) in an attempt to achieve comparability among various variables. However, studies still compare women and men with different levels of teaching ability. Thus, the meaning of the relationship between teaching evaluation and gender is no longer an *effect*, *impact*, or *bias* but just a *difference*.

The following hypothetical situation illustrates the limitation of interpreting a difference as evidence of *bias*. Consider a course evaluation questionnaire containing a few items activating a gender stereotype of *competence*. Consequently, students are more lenient toward male instructors because *men* are most likely to be seen as *competent*. Students are more stringent toward women because they are less likely to be seen as *competent*. Obtaining the same evaluation is harder for women than men due to an irrelevant aspect in the evaluation (the item content). A *biasing factors* study would show, for example, a trivial difference captured by a correlation or regression coefficient close to zero. Ignoring that item content creates an unfair teaching evaluation suggests that the trivial difference between women and men reflects an unbiased evaluation when the opposite is true. Given the item content, achieving the same result is harder for women than men. Thus, freeing the item from the biasing content would lead to a change in the difference in teaching evaluation between men and women (Rivera & Tilcsik, 2019).

Measurement bias and test fairness

Test fairness comprises theory and methods that help examine if a given measurement produces unfair differences between subgroups, in other words, if we employ the same or a different yardstick for men or women or by race or ethnicity of the participants. *Fairness* has been an enduring concern in educational measurement for decades in various settings, including personnel selection, standardized testing, college admission, and psychological testing (Camilli, 2006; Camilli, 2013; Zumbo, 1999). *Fairness* is a condition establishing the validity of a measure and the decisions the measure informs (AERA, APA, NCME, 2014).

A fairness study aims to "sort out whether the reasons for group differences are due to factors beyond the scope of the test […] or artifactual" (Camilli, 2006, p. 225). A fairness study contributes to "identify and remove construct-irrelevant barriers to maximal performance for any examinee" (AERA, NCME, APA, 2014, p. 190). There is a variety of tools for conducting fairness studies: "Such analyses could employ a range of methodologies, including those appropriate for small sample sizes, such as expert judgment, focus groups, and cognitive labs. Both qualitative and quantitative sources of evidence are important in evaluating whether items are psychometrically sound and appropriate for all relevant subgroups" (AERA, NCME, APA, 2014, p .193).

An essential aspect defining the use of the term *fairness* from an educational measurement standpoint is defining who benefits from the potential advantage. As a convention, the group with the social advantage is the *reference* group. The group with the disadvantage is the *focal group* (AERA, APA NCME, 2014). The decision of what individuals are part of the *focal* and *reference* groups relies upon social, political, and regulatory aspects (Camilli, 2006; Zumbo, 1999). For instance, historical and legally relevant focal groups in the US context include American Indian, Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian, Pacific Islander, and non-native English.

The second aspect of *fairness* relates to the definition of *bias*. From a legal perspective, *disparate impact* is the difference in performance between *reference* and *focal* groups (Camilli, 2006; Zumbo, 1999). This definition conveys that *disparate impact* does not imply *bias*. Bias refers specifically to systematic measurement error in test scores (AERA, APA, NCME, 2014). Systematic measurement error (or *bias*) produces deflation or inflation (additive error) or changes in the correlation coefficients with other variables (correlational error) (Viswanathan, 2005). An example of a bias affecting faculty teaching evaluation occurs when students respond unattentively (Bassett et al., 2017) or relying on a response style (Valencia, 2020). Bias (measurement error) produces a distorted picture of the instructors' level of teaching ability as a whole and inaccuracy in the correlation coefficients between teaching evaluation and other variables.

From a *fairness* perspective, bias relates to "construct-underrepresentation or construct-irrelevant components of test scores that differentially affect the performance of different groups of test-takers" (AERA, APA,, and NCME, 2014, p. 40). As in the previous example, item content triggers an irrelevant component in students' response (construct-irrelevant variance), the stereotype of competence, and creates a difference in the item's interpretation that depends on the instructor's gender. In this last specific sense, *bias* results in an arbitrary difference in performance between subgroups.

Statistical bias and differential item functioning

A typical step in a fairness study is the examination of *statistical bias*. Mathematical models help compare *focal* and *reference* groups with similar abilities. One approach involves identifying an external criterion that should correspond, in our case, to another already valid teaching ability measure. The goal is to determine that the prediction of the criterion using the teaching evaluation is equivalent between the focal group (e.g., women) and the reference group (e.g., men). Evidence of a differential prediction would reflect *statistical bias* because the expectation is the equivalence of the regression parameters (e.g., slope and intercept). Evidence of *statistical bias* suggests that the meaning of the faculty teaching evaluation is not the same in the *focal* and *reference* groups (Camilli, 2006).

Another strategy to collect evidence of *statistical bias* when no valid external criterion is available requires an internal criterion. The internal criterion strategy employs the same set of items to examine differences between *focal* and *reference* groups. The fundamental question that internal criterion strategies address is "Is this item measuring the same thing for two groups relative to the other items?" (Camilli, 2006, p. 229).

Differential item functioning (DIF) is an internal criterion strategy (uses the same items included in the teaching evaluation) that collect evidence of *statistical bias*. *Statistical bias* relates to the lack of equivalence of the attributes describing the functioning of an item between focal and reference groups (Camilli, 2006; Zumbo, 1999).

A relevant attribute characterizing the functioning of an item is its difficulty.[1] In traditional item analysis, the difficulty is the proportion of participants that succeed or endorsed high scores (Kline, 2005). For instance, obtaining a rating of 5 out of 5 in the item "the instructor arrived on time" seems easier to achieve than the same score of "5" in the item "the instructor created challenging lectures." The latter requires higher teaching ability than the former to deserve the same high score. Therefore, the difficulty of the second item should be the highest between the two items; thus, fewer instructors should receive higher scores compared to the first item. If an item triggers a gender stereotype, that item's difficulty should be higher for women than men.

DIF in faculty teaching evaluation

Institutions often devise methods for inferring teaching ability from questionnaires. Specifically, a faculty teaching evaluation involves using raters (i.e., the students) and multiple items using fixed response alternatives (e.g., Likert-type rating scale). With rating scale data, a typical method for obtaining an overall teaching ability measure is the sum (or average) across item responses or *total score* (Spector, 1992).

[1] A second attribute relates to item discrimination (the correlation between the item and the total score) and a third attribute relates to guessing (for measures with correct/incorrect answers). For parsimony, I address difficulty as the most relevant attribute explaining responses to items in a faculty teaching evaluation, but the same applies when focusing on discrimination.

Classical test theory (CTT) provides the rationale underlying the use of *total score*. CTT is a psychometric theory connecting *raw scores* and the *latent constructs* they target (Spector, 1992; Traub, 1997). CTT has strong assumptions about raw scores that are hard to comply with (Shavelson and Webb, 2006; Hambleton, Swaminathan, & Rogers, 1991). One limitation is the assumption that all items reflect teaching ability in the same way (e.g., "arriving on time" matters as much as "making the course intellectually stimulating"). In other words, CTT assumes that items' functioning is homogeneous. CTT only accounts for random error (a source of error detrimental to reliability), ignoring systematic error (or bias). Lastly, CTT and Likert-type scales are notorious for producing only ordinal data affecting a significant group of data analysis strategies utilized in faculty teaching evaluation (Boring et al., 2016; Stark & Freishtat, 2014). Thus, the *total score* is too simple to address the complexity of faculty teaching evaluation, examine fairness, and support personnel decisions based on these evaluations.

A helpful framework for overcoming the CTT limitations and examining DIF is the rating scale model (RSM) (Andrich, 1978; Wright & Masters, 1982). The RSM is a psychometric model appropriate for measures with response alternatives that are the same across all items (e.g., Likert-type scales). Prior research reports aspects of the validity of faculty teaching evaluations using the RSM (Haladyna & Hess, 1994; Meyer et al., 2017; Setari et al., 2016; Van Zile-Tamsen, 2017).

The RSM relates the probability of choosing a response option to a given item to three components: (1) the student's attitude toward the teaching ability of the instructor, (2) the difficulty of the item, and (3) the difficulty of choosing a specific response option. For instance, the RSM predicts that a student is more likely to select "not at all" for item "I found the course intellectually stimulating" if (1) she/he has a relatively low overall appraisal of teaching ability (low student's attitude), (2) the item expresses an attribute of teaching harder to accomplish (high item difficulty), and (3) the transition between adjacent response alternatives is harder (high threshold). The latter would occur if instead of selecting between "not at all" and "somewhat," the response alternatives were "not at all" and "moderately" since the width between "not at all" and "moderately" is larger than between "not at all" and "somewhat." Equation 1 formally presents the RSM (Bond & Fox, 2015, p. 350):

$$P_{nik} = \frac{e^{B_n - [D_i + F_k]}}{1 + e^{B_n - [D_i + F_k]}} \tag{1}$$

In Eq. 1, the probability of selecting the response option k in item i for student n depends on B_n representing the attitude of the student n, D_i that accounts for the difficulty of item i and F_k that "reflects the level at which the likelihood of being observed in a given response category (below the threshold) is exceeded by the likelihood of being observed in the next higher category" (Bond & Fox. p. 116.). The three components, the students' attitude, item difficulties, and thresholds, are unknown and are inferred from the raw responses following iterative estimation methods. The analysis produces estimates of the level of attitude for each student, a difficulty for each item, a fixed (the same across items) set of thresholds (one minus the number of response alternatives), and information regarding how well the data adheres to the RSM.

Equation 3 builds on Eq. 2 and captures gender differences because it includes the instructor's gender (G) as a facet that accounts for the probability of selecting a specific response alternative:

$$P_{nik} = \frac{e^{B_n - [D_i + F_k - G]}}{1 + e^{B_n - [D_i + F_k - G]}} \tag{2}$$

The DIF analysis compares item difficulties and thresholds between the *focal* and reference *groups* (Camilli, 2006). Adapting the previous equation to a DIF analysis results in Eqs. 3 and 4:

$$P_{gnik} = \frac{e^{B_{gn} - \left[D_{gi} + F_k - G\right]}}{1 + e^{B_{gn} - \left[D_{gi} + F_k - G\right]}} \tag{3}$$

$$P_{gnik} = \frac{e^{B_{gn} - \left[D_{gi} + F_{gk} - G\right]}}{1 + e^{B_{gn} - \left[D_{gi} + F_{gk} - G\right]}} \tag{4}$$

P_{gnik} defines the probability of person n from subgroup g (*reference* and *focal*, r and f, respectively) of choosing response option k to item i. In Eq. 3, only difficulties may vary by group (D_{gi}), while in Eq. 4, both difficulty (D_{gi}) and threshold (F_{gk}) may vary by group. A difference between the *focal* and *reference* group in either difficulty ($B_{fi} \neq B_{ri}$, Eqs. 3 and 4) or thresholds ($F_{fk} \neq F_{rk}$, Eq. 4) would suggest DIF.

This study

A difference in teaching ability between women and men is not evidence of a gender-biased teaching evaluation. Implementing an experimental study to capture the effect of gender is unfeasible in most settings. Under the conditions above, how can practitioners and administrators examine the fairness of a faculty teaching evaluation using routinely collected data?

This study aims to illustrate the use of a differential item functioning analysis (DIF) and provide evidence of the fairness of a teaching evaluation. The study helps answer the following two questions. First, are individual items in the faculty teaching evaluation measuring the same thing for female and male instructors relative to the other items? Second, what is the size of the *disparate impact* or gender gap, if any?

Methods

Participants

The study examined students' responses to a teaching evaluation questionnaire administered to education graduate students at a large university in North America. The university granted access to data containing only students' raw responses, instructor's gender information, type of degree program (Master or Ph.D.), and a variable identifying two academic departments (departments A and B) due to confidentiality concerns. Two other academic departments in the institution employed a different questionnaire and are not part of the study. Thus, the data contains no other student, instructor, course, or program information preventing further analysis. Students received an email requesting voluntary participation. The mode of administration of the questionnaire was online. Valencia (2020) reported a complete description of the administration procedure in the context of a separate construct validation study. The present study focuses on a subsample from the original data collected in winter 2016. In the two departments, 70% of the students were enrolled in Master and the rest in Ph.D. programs and almost a third of the questionnaires rated male instructors (reference group) while two-thirds rated female instructors (focal group). The total number of students with complete information for the analysis is 514 in department A and 418 in department B.

Teaching evaluation questionnaire

The results of the teaching evaluation inform both formative and summative decisions for faculty and program and curriculum improvement in this institution. The online questionnaire contains eight items with statements targeting the instructor's teaching ability. The eight statements are "I found the course intellectually stimulating" (item 1), "The course provided me with a deeper understanding of the subject matter" (item 2), "The instructor created a course atmosphere that was conducive to my learning" (item 3), "Course projects, assignments, tests, and/or exams improved my understanding of the course material" (item 4), "Course projects, assignments, tests, and/or exams provided opportunity for me to demonstrate an understanding of the course material" (item 5), "The instructor explained the learning objectives for the course" (item 6), "The course instructor demonstrated respect for diversity (e.g., race, gender, ability, religion, sexual orientation, etc.) in the classroom" (item 7), and "The course instructor encouraged students to express their own ideas in the class" (item 8). Instructions requested students to rate the level of agreement with each statement using one of the following five response alternatives: "not at all," "somewhat," "moderately," "mostly," and "a great deal." Students answered the same questionnaire regardless of the department, course, or degree program.

Data analysis

Previous equations sustain four quantitative analyses (models 1 to 4). The first analysis involves estimating students' attitudes, item difficulties, and thresholds (model 1). The analysis checks the overall quality of the SET data, the extent to which the data complies with the psychometric model depicted in Eq. 1, and is useful for measuring teaching ability. Two indexes for ascertaining data quality are the inlier-sensitive (or information-weighted) mean square and outlier sensitive (or information-weighted) mean square (infit and outfit, henceforth). These indexes summarize the amount of residual information, the difference between expected and predicted values following the model. Infit or outfit excessively below 1 indicates less variation than expected by the model (i.e., measurement redundancy). Infit or outfit excessively above 1 indicates more variation than expected by the model (i.e., there is construct-irrelevant variance). A general rule of thumb recommends marking values under 0.5 or over 1.5 as problematic (Osteen, 2010; Wright & Linacre, 1994; Wu et al., 2007). Another guideline suggests providing greater importance to infit than outfit because the second is very sensitive to large individual residuals (Ames & Penfield, 2015; de Ayala, 2009).

The following step targets gender differences and involves fitting the data to Eqs. 2, 3, and 4. Equation 2 adds an instructor's gender as a measurement facet (model 2). Equation 3 adds gender effect and item difficulties by gender (model 3). Equation 4 adds gender effect and both item difficulties and thresholds varying by gender (model 4). Evidence of DIF derives from comparing the overall fit of model 2 and model 3. If model 2 offers a better fit than model 3, then the functioning of item difficulties is equivalent between reference (men) and focal (women) groups. On the contrary, a model 3 with a better fit than model 2 provides evidence of DIF. The same applies to the examination of differential thresholds by gender (model 4).

Deviance is a goodness of fit index that allows relative comparisons between competing models, with smaller deviances indicating a better relative fit. A rule of thumb suggests interpreting a difference in deviance between zero a two as a "substantial" level of empirical support, between four and seven as "considerably less" level of empirical support, and above ten as "essentially none" level of empirical support for a model (Burnham & Anderson, 2002, p. 71). Additionally, a likelihood ratio χ^2 test of the difference in deviance (with the difference in parameters as the degree of freedoms) examines the hypothesis of equal model fit. A statistically significant χ^2 indicates that the competing model shows a statistically significant better fit.

If the analysis reveals DIF due to difficulty (or thresholds) varying by gender, the final step examines each individual item's difficulty/threshold by gender. Estimates by gender are expressed as a difference from the average item difficulty or threshold, respectively. A standardized value (the estimate divided by the standard error) above |2| suggests a statistical bias for the specific item or threshold (Bond & Fox, 2015; Camilli, 2006; Wu et al., 2007). All models were fitted using the software *Conquest* 5 for the *macOS* platform (Adams et al., 2020).

Results

Model 1 (SET quality)

The analysis produced seven-item difficulties, three thresholds, a mean, and a variance describing the students' attitude level about teaching. The estimation converged quickly and satisfactorily after 34 iterations for department A and after 70 iterations for department B. The distribution of students' attitudes was negatively skewed in the two departments, with more students leaning toward reporting high teaching ability. Students' attitude extended from −2 logits to +10 logits, with an average teaching ability of 2.99 logits (*SD* = 2.45) in department A and 4.4 logits (*SD* = 2.46) in department B. The reliability in the two departments is adequate, with a person-separation index of 0.86 in department A and 0.80 in department B. Reliability under CTT using Cronbach's alpha internal consistency coefficient is spuriously much higher, with values of 0.95 and 0.94 for each department, respectively.

Table 1 presents item difficulties, thresholds, and fit indexes (infit and outfit). Item difficulties ranged in the ±2 logit intervals. Other educational measures span over similar intervals, and between ±4 logits or more (see Bond & Fox, 2015; Wright & Masters, 1982; Wu et al., 2007 for applied examples). Thus, the questionnaire could benefit from more items targeting harder-to-achieve teaching attributes. Item 7, "The course instructor demonstrated respect for diversity in the classroom," is the most aggregable statement in the two departments, exemplifying an easy-to-achieve attribute of teaching. The most difficult item to endorse in department A is item 3 "The instructor created a course atmosphere that was conducive to my learning." Item 1, "I found the course intellectually stimulating," is the most difficult item in department B.

Table 1 Item difficulties, thresholds, and residual-based fit indexes for model 1

	Department A				Department B			
Parameter	Est.	SE	Outfit	Infit	Est.	Error	Outfit	Infit
Item 1	0.64	0.07	0.85	0.86	0.73	0.09	0.94	0.93
Item 2	0.28	0.07	1.06	0.84	0.27	0.09	0.81	0.94
Item 3	0.72	0.07	0.78	1.01	0.56	0.09	0.86	1.04
Item 4	0.31	0.07	0.74	0.88	0.57	0.09	0.79	0.85
Item 5	0.22	0.07	0.69	0.79	0.38	0.09	0.68	0.90
Item 6	−0.14	0.07	1.07	1.21	0.27	0.09	0.94	0.99
Item 7	−1.28	0.09	1.54	1.74	−1.59	0.14	1.18	1.68
Item 8[a]	−0.76	0.08	1.00	1.22	−1.20	0.12	0.79	1.33
Threshold 1	−2.40	0.14	4.62	1.50	−3.31	0.33	5.67	1.49
Threshold 2	−0.35	0.09	1.10	1.67	−0.22	0.16	0.92	1.67
Threshold 3	0.48	0.08	1.29	1.54	0.67	0.14	1.26	1.58
Threshold 4[a]	2.27		1.30	1.22	2.86		1.15	1.25

[a] Fixed estimates for model identification purposes

Est., difficulty/threshold estimate; *SE*, standard error of measurement; *outfit*, outlier-sensitive mean square fit; *infit*, information-weighted mean square fit

The fit of the data to the RSM seems adequate in the two departments because most infit and outfit are around their expected value of 1 and within the .5 and 1.5 rule of thumb range. There is no problematic redundancy because all infit and outfit are above the 0.5 lower bound cutoff. The thresholds 1 to 4 are ordered from negative to positive values following how the response alternative should function: Selecting higher response alternatives should require more attitude level. Outfit for threshold 1 (selecting "not at all" versus "somewhat") exceeds the upper bound 1.5 cutoff. However, the infit falls within the expected value. Infit for thresholds 2 and 3 are slightly above the 1.50 cutoff, suggesting some issues in the response scale utilization. Despite these three areas for further improvement (for instance, evaluating the number and labels for the response scale), the results show SET data reliable and useful for gender analysis. Results also suggest slight differences in the way the questionnaire worked between departments, in specific, what were the hardest items.

Differences by gender and DIF

Table 2 presents summary information for four analyses following Eqs. 1, 2, 3, and 4. The first noteworthy result relates to the improvement in the goodness of fit (deviance) of all three competing models (models 2, 3, and 4) compared to model 1 in the two departments. Results reveal a large (above 10) and statistically significant difference in deviance from model 1 to the gender difference analysis (Eq. 2) in department A ($\chi^2(1, N = 514) = 56.69$, $p < 0.001$), offering null empirical support for model 1. The results of fitting Eq. 2 reveals that women receive 1.75 logits more favorable attitude toward teaching ability, representing 0.75 standard deviations or a medium-size effect (Cohen, 1988; Ellis, 2010). The difference by gender is statistically significant (χ^2(1, $N = 514) = 60.06$, $p < .001$). In contrast, the change in deviance in department B offers empirical support for model 1 rather than the *gender difference* model. The change in deviance is nonstatistically significant ($\chi^2(1, N = 418) = 0.59$, $p = 0.39$). Coherently, the

Table 2 Comparison among four models for examining DIF by instructor's gender in SET data from two academic departments

Equation/model	(1)	(2)	(3)	(4)
Department A				
Model deviance	7113,42	7057,73	6990,87	6986,19
Parameters	7137,42	7083,73	7030,87	7032,19
Deviance difference	34	104	104	100
Degrees of freedom	12	13	20	23
LR *p*-value		55,69	66,86	4,69
Department B				
Model deviance	4537,37	4536,78	4531,79	4528,18
Parameters	4561,37	4562,78	4571,79	4574,18
Deviance change	70	81	80	124
Degrees of freedom	12	13	20	23
LR *p*-value		0,59	4,99	3,62

(1), baseline model; (2), gender difference model; (3), differential item difficulty model; (4), differential difficulty and threshold model; *LR*, likelihood ratio

difference of .224 logits favoring men represents about .09 of the standard deviation, a trivial effect size not statistically significant ($\chi^2(1, N = 418) = 0,\ 51,\ p = .43$).[2]

The first DIF analysis includes separate item difficulty for women and men (3). The model is effective in reducing deviance from model 2 in the two departments. The deviance change is statistically significant in department A ($\chi^2(7, N = 514) = 66.86,\ p < 0.001$), providing no empirical support to model 2. On the contrary, the magnitude of the deviance change in department B is too small and not statistically significant ($\chi^2(7, N = 418) = 4.99, p = .12$). The results suggest a systematic interaction between gender and items coherent with DIF only in department A.

The second DIF analysis includes specific item difficulty and thresholds by gender (model 4). The change in deviance from model 3 to model 4 suggests a substantial level of empirical support for the more simply model 2. The likelihood ratio test suggests a nonstatistically significant difference in deviance in department A ($\chi^2(3, N = 514) = 4.69, p = .08$) and department B ($\chi^2(3, N = 418) = 3.62,\ p = .12$). Thus, the results suggest no interaction between gender and thresholds. Accordingly, students did not employ the response scale differently depending on gender, and DIF relates mainly to the interaction between the instructor's gender and items exclusively in department A.

Examining DIF in department A

Table 3 shows differences in the functioning of item difficulty by gender in department A. Estimates are a deviation from the average item difficulty when the instructors are men (top half) or women (bottom half). Values represent how much difficult it is for women or men to receive the same score in the item. The bottom part of the table only differs in sign from the upper part. Table 3 also shows difficulty divided by the standard error to determine what specific item exhibits DIF (labeled Z in the last column in Table 3). As a rule of thumb, values of Z above 2 flag an item with DIF.

[2] With a CTT approach, the conclusion is similar, but the size of the difference is smaller, 0.61 and 0.4 standard deviations for departments A and B, respectively.

Table 3 Department A estimates of item difficulty by instructor's gender (model 3)

Item	Est.	SE	Outfit	Infit	Z
Reference group: men					
1. I found the course intellectually stimulating	0.17	0.07	0.95	0.87	2.41
2. The course provided me with a deeper understanding of the subject matter	0.21	0.07	0.87	0.88	2.96
3. The instructor created a course atmosphere that was conducive to my learning	0.21	0.07	0.85	0.94	3.03
4. Course projects, assignments, tests, and/or exams improved my understanding of the course material	−0.12	0.07	0.91	0.91	−1.70
5. Course projects, assignments, tests, and/or exams provided opportunity for me to demonstrate an understanding of the course material	−0.14	0.07	0.93	0.87	−1.96
6. The instructor explained the learning objectives for the course	0.30	0.07	1.32	1.37	4.07
7. The course instructor demonstrated respect for diversity (e.g., race, gender, ability, religion, sexual orientation, etc.) in the classroom	−0.48	0.09	2.08	1.63	−5.30
8. The course instructor encouraged students to express their own ideas in the class[a]	−0.15	0.08	0.87	1.16	−1.85
Focal group: women[a]					
1. I found the course intellectually stimulating	−0.17	0.07	0.95	0.92	−2.41
2. The course provided me with a deeper understanding of the subject matter	−0.21	0.07	0.75	0.85	−2.96
3. The instructor created a course atmosphere that was conducive to my learning	−0.21	0.07	0.78	1.08	−3.03
4. Course projects, assignments, tests, and/or exams improved my understanding of the course material	0.12	0.07	0.72	0.93	1.70
5. Course projects, assignments, tests, and/or exams provided opportunity for me to demonstrate an understanding of the course material	0.14	0.07	0.69	0.78	1.96
6. The instructor explained the learning objectives for the course	−0.30	0.07	1.05	1.06	−4.07
7. The course instructor demonstrated respect for diversity (e.g., race, gender, ability, religion, sexual orientation, etc.) in the classroom	0.48	0.09	1.3	1.62	5.30
8. The course instructor encouraged students to express their own ideas in the class	0.15	0.08	1.14	1.32	1.85

Est., difficulty estimate; *SE*, standard error of measurement; *outfit*, outlier-sensitive mean square fit; *infit*, information-weighted mean square fit; *Z*, estimate/standard error

[a] Fixed estimates for model identification purposes

Examination of Table 3 suggests three types of items: items with no evidence of DIF by gender (three items), items easier to endorse for women (four items), and an item easier to endorse for men.

Items with no DIF are item 4 ("Course projects, assignments, tests, and/or exams improved my understanding of the course material"), item 5 ("Course projects, assignments, tests, and/or exams provided opportunity for me to demonstrate an understanding of the course material"), and item 8 ("The course instructor encouraged students to express their own ideas in the class."). Items that are notably easier to endorse about the teaching of women are item 1 ("I found the course intellectually stimulating"), item 2 ("The course provided me with a deeper understanding of the subject matter"), item 3 ("The instructor created a course atmosphere that was conducive to my learning"), and item 6 ("The instructor explained the learning objectives for the course)." The two first of these items are also the hardest ones to endorse, and as a group, these items contain or refer to teaching effectiveness. The only item favoring men is item 7 ("The course instructor demonstrated respect for diversity in the classroom"). This item captures the easiest attribute of teaching to endorse and is the only item not directly targeting an aspect of teaching effectiveness. The size of the difference favoring men in this item (−0.48 logits easier) is sizable and accounts for 0.21SD.

Conclusion

The study illustrated a *differential item functioning* (DIF) analysis that cast light on two critical issues: whether a teaching evaluation measured the same thing for women and men and the size of the gender *disparate impact*, if any. The teaching evaluation comprised an online questionnaire answered by education graduate students from two academic departments. The analysis indicated no DIF in one of the departments, suggesting that these items were free from statistical bias and measured teaching in the same fashion for women and men. The findings help support the fairness of the evaluation allowing meaningful comparison of teaching ability by gender. The findings also suggested no *disparate impact* affecting women in this department. The same kind of analysis revealed *DIF* in five items in the second department. Four items were harder for men than women with comparable levels of teaching ability. The analysis also showed that women received a higher teaching evaluation.

These findings illustrate the value of two critical aspects from a test fairness perspective. The first critical aspect is the distinction between *disparate impact* and *bias*. The distinction enables the study of test fairness without the misconception of interpreting a *difference* as evidence of a biased evaluation (Feldman, 1993; MacNell et al., 2015; Marsh, 1987; Theall & Franklin, 2001; Wachtel, 1998). A *biasing factors* study design may well (1) report a null difference that reflects an actual difference between subgroups or (2) hide the actual difference because of an unfair teaching evaluation method. An example is the results from department A. There is uncertainty about how much larger the difference in favor of women would have been had the group of items showed no statistical bias. Thus, researchers, practitioners, and administrators should verify that the teaching evaluation contains no substantial statistical bias before examining *disparate impacts*.

Second, the definition of the *focal* and *reference* groups forces us to identify the groups under scrutiny and justify why the focal group needs protection (Camilli, 2006, 2013; Zumbo, 1999). History, theory, and empirical evidence encourage studying the fairness of teaching evaluation for women and men. In opposition, the *biasing factors* are often given by the variables already available in SET data (MacNell et al., 2015) without supporting the theory (Marsh, 1987). Defining the groups ahead of the fairness study prevents spuriously statistically significant findings (phishing) because of data-driven analysis of the many potential *biasing factors*. Thus, administrators should examine and document the fairness of teaching evaluation for all culturally relevant underrepresented subgroups by gender, race, or ethnicity before using the evaluation for informing personnel and other high-stake decisions. Current gender inequalities call for adopting the best available evaluation practices, especially if one considers the adverse effects of COVID-19 on teaching.

There is a reasonable apprehension about a reduction of transparency with the introduction of novel teaching evaluation practices such as the rating scale model (RSM) proposed here and elsewhere (Haladyna & Hess, 1994; Meyer et al., 2017; Setari et al., 2016; Van Zile-Tamsen, 2017). First, reduced transparency should not be confused with higher complexity. Current practices are also quantitative and anchored in measurement theory (classical test theory). Although they seem familiar, there is already a problem with the transparency of current practices related to a general lack of understanding and misconceptions affecting teaching evaluation (Theall & Franklin, 2001; Penny, 2003; Boysen, 2015, Boysen et al., 2014). The ultimate lack of transparency relates to not examining test fairness in faculty evaluation or using unfit practices. The utilization of defensible methods and proper documentation contributes to monitor and minimize the gender gap, increasing transparency and trust in the long run. In this regard, tools for examining fairness are readily available. Fairness captures a whole chapter of the latest *Standards for Educational and Psychological Measurement* (AERA, APA, and NCME, 2014). DIF and similar analyses are routinely conducted in various educational settings, including large-scale assessment, college admission tests, and personnel selection. Several licensed and free software for quantitative analysis already includes DIF routines. Institutions may improve their competencies on educational measurement in different ways, for instance, through training or by recruiting qualified staff.

Examining DIF using the RSM is convenient for summated rating scales (e.g., Likert scales), the foundation of most course evaluation questionnaires. Though, the strategy is only one of the many already available. Other DIF internal methods cover measures featuring correct/incorrect scoring, rubrics with a different number of performance levels (e.g., classroom observation protocols, portfolios), and measures where discrimination, along with difficulty, is an essential attribute of item responses. Other strategies also include the use of external criteria (i.e., differential test prediction) suitable for the rare situation where a valid measure of teaching ability is available. There are different types of DIF (uniform and non-nonuniform) not covered in this study. More importantly, strategies for examining fairness reach beyond the realm of statistics to include qualitative evidence and logical argumentation (Camilli, 2013). The study of fairness in teaching evaluation and other faculty evaluation methods may well begin during the development process (e.g., Educational Testing Service, 2016). Thus, instead of building the case for a unique strategy, logical reasoning, along with the collection of diverse types of evidence, is the most appropriate course of action for addressing gender bias in faculty evaluation. The challenge for administrators is to find a way to elaborate, test, and demonstrate the neutrality of the content and procedure so that, for instance, the intellectually stimulating teaching from women gets the same evaluation as the comparable intellectually stimulating teaching performed by men.

Limitations and recommendations

The concept of test fairness represents a step forward from the *biasing factors* literature. However, no perspective is flawless, and these tools are limited in several ways. A significant limitation relates to sample size. DIF analysis requires information from various instructors and students to inform whether a bias is systematic. The tool allows inferences about the group but virtually no insight into the fairness of one specific instructor (Camilli, 2013). Another limitation relates to the reasons underlying DIF. The tools on their own do not provide explanations for *DIF* nor *disparate impact*. For instance, arguing that items about “learning” trigger gender stereotypes favoring women is disputable. Thus, after an item shows DIF, there is still room for deliberation on whether DIF threatens the evaluation’s fairness. Also, the idea of using the same yardstick could be potentially unfair if the teaching conditions are already harder for some instructors than others. The fact that specific groups of instructors often teach larger, entry-level courses may affect their opportunity to show some of their teaching skills. For instance, the same teaching evaluation may not capture relevant differences between the novel and experienced teachers. DIF and other tools for examining fairness are limited to detect the extent to which there is statistical bias, requiring other types of evidence and reasoning to put any gender, racial, and ethnic difference into context.

Drawing from the K-12 teaching evaluation literature (Berliner, 2005), the evaluation of teaching in higher education also seems difficult and context-dependent. The generalizability of the gender bias findings from this study is scarce. The same analysis in two academic departments resulted in different conclusions. The comparison between departments illustrates how generalization about gender differences may lead to confusion. Theory and evidence suggest complex ways in which gender and other stereotypes can affect teaching evaluations (Basow & Martin, 2012; Laird et al., 2011). The characteristics of the evaluation method, including the dimensions of teaching ability, item content, participants, response format, and settings, can trigger changes in the evaluation results (e.g., Rivera & Tilcsik, 2019; Zipser & Mincieli, 2018), affecting validity (Messick, 1995). As a result, fairness and measurement validity are local instead of a property of the evaluation method (AERA, NCME, APA, 2014; Bond & Fox, 2015).

Lastly, the study brings attention to the ubiquitous use of SET. Likert-type scales provide a "quick and easy way of producing some sort of overall score" (Bond & Fox, p. 112). However, moving away from SET (as the ASA statement suggests) requires demonstrating that the alternative overpasses the quality and benefits of the current method. A way of coping with the limitations of Likert-type scales and students' biases is by building a persuasive argument about validity based on various forms of evidence and argumentation, including fairness (AERA, APA, NCME, 2014; Zumbo, 1999). Thus, an area of improvement in teaching evaluation relates to the scarce documentation of the validity of in-house measures available to their faculty and communities. With this documentation, administrators can better argue about maintaining, improving or replacing current practices, adding more transparency and credibility to a highly controversial topic. There is no bias-free method of faculty teaching evaluation. Thus, a recommendation is to continuously examine and improve our current and new practices over time, strengthening the organizational structures to do so and opening the information to rigorous external scrutiny while taking precautions to maintain the confidentiality of the information.

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2020). *ConQuest: Generalised item response modelling software (4.5.2) [Computer Software]*. Australian Council for Educational Research.

Alhija, F. (2017). Guest editor introduction to the special issue "contemporary evaluation of teaching: Challenges and promises". *Studies in Educational Evaluation, 54*(Supplement C), 1–3. https://doi.org/10.1016/j.stueduc.2017.02.002.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Sociological Association. (2019). *Reconsidering student evaluations of teaching*. American Sociological Association. Retrieved November 6, 2019, from https://www.asanet.org/press-center/press-releases/reconsidering-student-evaluations-teaching

Ames, A. J., & Penfield, R. D. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice, 34*(3), 39–48. https://doi.org/10.1111/emip.12067.

Andersen, K., & Miller, E. D. (1997). Gender and Student Evaluations of Teaching. PS: *Political Science and Politics, 30*(2), 216. https://doi.org/10.2307/420499.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573. https://doi.org/10.1007/BF02293814.

Arbuckle, J., & Williams, B. D. (2003). Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles, 49*(9–10), 507–516. https://doi.org/10.1023/A:1025832707002.

Basow, S. A., & Martin, J. L. (2012). Bias in student evaluations. In *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 40–49). Society for the Teaching of Psychology.

Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education, 18*(2), 91–106. https://doi.org/10.1007/s11092-006-9001-8.

Bassett, J., Cleveland, A., Acorn, D., Nix, M., & Snyder, T. (2017). Are they paying attention? Students' lack of motivation and attention potentially threaten the utility of course evaluations. *Assessment & Evaluation in Higher Education, 42*(3), 431–442. https://doi.org/10.1080/02602938.2015.1119801.

Bavishi, A., Madera, J. M., & Hebl, M. R. (2010). The effect of professor ethnicity and gender on student evaluations: Judged before met. *Journal of Diversity in Higher Education, 3*(4), 245–256. https://doi.org/10.1037/a0020763.

Bertrand, M. (2017). The glass ceiling. *Becker Friedman Institute for Research in Economics Working Paper No. 2018-38,* https://doi.org/10.2139/ssrn.3191467

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental measurement in the human sciences, third edition* (3rd ed.). Routledge.

Bonitz, V. S. (2011). *Student evaluation of teaching: Individual differences and bias effects*. Graduate Theses and Dissertations. 12211. Retrieved November 6, 2019, from https://lib.dr.iastate.edu/etd/12211

Boring, A., Ottoboni, K., & Stark, P. B. (2016). *Student evaluations of teaching (mostly) do not measure teaching effectiveness* (pp. 1–11). ScienceOpen Research. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1.

Boysen, G. A. (2015). Significant interpretation of small mean differences in student evaluations of teaching despite explicit warning to avoid overinterpretation. *Scholarship of Teaching and Learning in Psychology, 1* (2), 150–162. https://doi.org/10.1037/stl0000017.

Boysen, G. A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2014). The (mis)interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education, 39*(6), 641–656. https://doi.org/10.1080/02602938.2013.860950.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer-Verlag http://www.springer.com/gp/book/9780387953649.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (Fourth ed., pp. 221–256). Praeger Publishers.

Camilli, G. (2013). Ongoing issues in test fairness. *Educational Research and Evaluation, 19*(2–3), 104–120. https://doi.org/10.1080/13803611.2013.767602.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44*(5), 495–518. https://doi.org/10.1023/A:1025492407752.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

Cundiff, J. L., Danube, C. L., Zawadzki, M. J., & Shields, S. A. (2018). Testing an intervention for recognizing and reporting subtle gender bias in promotion and tenure decisions. *The Journal of Higher Education, 89*(5), 611–636. https://doi.org/10.1080/00221546.2018.1437665.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

Feldman, K. A. (1993). College students' views of male and female college teachers: Part II—Evidence from students' evaluations of their classroom teachers. *Research in Higher Education, 34*(2), 151–211. https://doi.org/10.1007/BF00992161

Gómez Cama, M., Larrán, M. J., & Andrades Peña, F. J. (2016). Gender differences between faculty members in higher education: A literature review of selected higher education journals. *Educational Research Review, 18*, 58–69. https://doi.org/10.1016/j.edurev.2016.03.001.

Haladyna, T., & Hess, R. K. (1994). The detection and correction of bias in student ratings of instruction. *Research in Higher Education, 35*(6), 669–687. https://doi.org/10.1007/BF02497081.

Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation*. Sage Publications.

Laird, T. F., Garver, A. K., & Niskodé-Dossett, A. S. (2011). Gender gaps in collegiate teaching style: Variations by course characteristics. *Research in Higher Education, 52*(3), 261–277. https://doi.org/10.1007/s11162-010-9193-0.

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40*(4), 291–303. https://doi.org/10.1007/s10755-014-9313-4.

Malisch, J. L., Harris, B. N., Sherrer, S. M., Lewis, K. A., Shepherd, S. L., McCarthy, P. C., Spott, J. L., Karam, E. P., Moustaid-Moussa, N., Calarco, J. M., Ramalingam, L., Talley, A. E., Cañas-Carrell, J. E., Ardon-Dryer, K., Weiser, D. A., Bernal, X. E., & Deitloff, J. (2020). Opinion: In the wake of COVID-19, academia needs new solutions to ensure gender equity. *Proceedings of the National Academy of Sciences, 117*(27), 15378–15381. https://doi.org/10.1073/pnas.2010636117.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*(3), 253–388. https://doi.org/10.1016/0883-0355(87)90001-2.

McClain, L., Gulbis, A., & Hays, D. (2017). Honesty on student evaluations of teaching: Effectiveness, purpose, and timing matter! *Assessment & Evaluation in Higher Education, 43*, 1–17. https://doi.org/10.1080/02602938.2017.1350828.

McPherson, M. A., & Jewell, R. T. (2007). Leveling the playing field: should student evaluation scores be adjusted? *Social Science Quarterly, 88*(3), 868–881. https://doi.org/10.1111/j.1540-6237.2007.00487.x.

McPherson, M. A., Jewell, R. T., & Kim, M. (2009). What determines student evaluation scores? A random effects analysis of undergraduate economics classes. *Eastern Economic Journal, 35*(1), 37–51.

Meyer, J. P., Doromal, J. B., Wei, X., & Zhu, S. (2017). A criterion-referenced approach to student ratings of instruction. *Research in Higher Education, 58*(5), 545–567. https://doi.org/10.1007/s11162-016-9437-8.

Onwuegbuzie, A. J., Daniel, L. G., & Collins, K. M. T. (2009). A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality & Quantity, 43*(2), 197–209. https://doi.org/10.1007/s11135-007-9112-4.

Osteen, P. (2010). An introduction to using multidimensional item response theory to assess latent factor structures. *Journal of the Society for Social Work and Research, 1*(2), 66–82. https://doi.org/10.5243/jsswr.2010.6.

Rivera, L. A., & Tilcsik, A. (2019). Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation. *American Sociological Review, 84*(2), 248–274. https://doi.org/10.1177/0003122419833601.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs*. American Educational Research Association.

Setari, A. P., Lee, J., & Bradley, K. D. (2016). A psychometric approach to the validation of a student evaluation of teaching instrument. *Studies in Educational Evaluation, 51*, 77–87. https://doi.org/10.1016/j.stueduc.2016.09.006.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., & Miller, V. D. (2007). The influence of student sex and instructor sex on student ratings of instructors: Results from a college of communication. *Women's Studies in Communication, 30*(1), 64–77. https://doi.org/10.1080/07491409.2007.10162505.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching the state of the art. *Review of Educational Research, 83*(4), 598–642. https://doi.org/10.3102/0034654313496870.

Stark, P., & Freishtat, R. (2014). *An evaluation of course evaluations*. ScienceOpen Research https://www.scienceopen.com/document/id/ad8a9ac9-8c60-432a-ba20-4402a2a38df4.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *New Directions for Institutional Research, 2001*(109), 45–56. https://doi.org/10.1002/ir.3.

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007.

Van Zile-Tamsen, C. (2017). Using Rasch analysis to inform rating scale development. *Research in Higher Education, 58*(8), 922–933. https://doi.org/10.1007/s11162-017-9448-0.

Valencia, E. (2020). Acquiescence, instructor's gender bias and validity of student evaluation of teaching. *Assessment & Evaluation in Higher Education, 45*(4), 483–495. https://doi.org/10.1080/02602938.2019.1666085.

Viswanathan, M. (2005). *Measurement Error and Research Design*. SAGE Publications Inc.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23*(2), 191–212. https://doi.org/10.1080/0260293980230207.

Weisshaar, K. (2017). Publish and perish? An assessment of gender gaps in promotion to tenure in academia. *Social Forces, 96*(2), 529–560. https://doi.org/10.1093/sf/sox052.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8* (3), 370.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Mesa Press.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0: Generalised Item Response Modelling Software*. ACER Press.

Zipser, N., & Mincieli, L. (2018). Administrative and structural changes in student evaluations of teaching and their effects on overall instructor scores. *Assessment & Evaluation in Higher Education, 43*(6), 995–1008. https://doi.org/10.1080/02602938.2018.1425368.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Directorate of

Human Resources Research and Evaluation, Department of National Defense. Retrieved November 6, 2019, from http://faculty.educ.ubc.ca/zumbo/DIF/handbook.pdf

Anderson, K. J., & Smith, G. (2005). Students' Preconceptions of Professors: Benefits and Barriers According to Ethnicity and Gender. *Hispanic Journal of Behavioral Sciences, 27*(2), 184–201. https://doi.org/10.1177/0739986304273707.

Aleamoni, L. M., & Hexner, P. Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. Instructional Science, 9(1), 67–84. https://doi.org/10.1007/BF00118969.

Educational Testing Service. (2016). *ETS international principles for the fairness of assessments*. Princeton, NJ: Author; Berliner, 2005.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Sage Publications, Inc.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity,bias, and utility. *American Psychologist, 52*(11), 1187–1197. https://doi.org/10.1037/0003-066X.52.11.1187.

Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender Bias in Teaching Evaluations. *Journal of the European Economic Association, 17*(2), 535–566. https://doi.org/10.1093/jeea/jvx057.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741.

Ory, J. C. (2001). Faculty Thoughts and Concerns About Student Ratings. New Directions for Teaching and Learning, 2001(87), 3–15. https://doi.org/10.1002/tl.23; American Sociological Association. (2019, September 9). Reconsidering Student Evaluations of Teaching. American Sociological Association. https://www.asanet.org/presscenter/press-releases/reconsidering-student-evaluations-teaching

Penny, A. R. (2003). Changing the Agenda for Research into Students' Views about University Teaching: Four shortcomings of SRT research. *Teaching in Higher Education, 8*(3), 399–411. https://doi.org/10.1080/13562510309396.

Spector, P. E. (1992). *Summated rating scale construction: An introduction*. Newbury Park, CA: Sage Publications.

Traub, R. E. (1997). Classical Test Theory in Historical Perspective. *Educational Measurement: Issues and Practice, 16*(4), 8–14. https://doi.org/10.1111/j.1745-3992.1997.tb00603.x.

Wagner, N., Rieger, M., & Voorvelt, K. (2016). Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of Education Review, 54*, 79–94. https://doi.org/10.1016/j.econedurev.2016.06.004.

Shavelson, R. J., & Noreen, W. (2006). Generalizability Theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of Complementary Methods in Education Research* (pp. 309–322). Washington DC: Lawrence Elbraum Associates, Inc.

Mitchell, K. M. W., & Martin, J. (2018). Gender Bias in Student Evaluations. PS: *Political Science & Politics, 51* (03), 648–652. https://doi.org/10.1017/S104909651800001X.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.