Bias in Student Ratings of Instruction: A Systematic Review of Research from 2012 to 2021

Brenda M. Stoesz, Amy E. De Jaeger, Matthew Quesnel,
Dimple Bhojwani, & Ryan Los
Centre for the Advancement of Teaching and Learning,
University of Manitoba

Abstract

Student ratings of instruction (SRI) are commonly used to evaluate courses and teaching in higher education. Much debate about their validity in evaluating teaching exists, which is due to concerns of bias by factors unrelated to teaching quality (Spooren et al., 2013). Our objective was to identify peer-reviewed original research published in English from January 1, 2012, to March 10, 2021, on potential sources of bias in SRIs. Our systematic review of 63 articles demonstrated strong support for the continued existence of gender bias, favoring male instructors and bias against faculty with minority ethnic and cultural backgrounds. These and other biases must be considered when implementing SRIs and reviewing results. Critical practices for reducing bias when using SRIs include implementing bias awareness training and avoiding use of SRIs as a singular measure of teaching quality when making decisions for teaching development or hiring and promotion.

Keywords: gender bias, postsecondary education, student evaluation of teaching (SET), teacher evaluations

Bias in Student Ratings of Instruction: A Systematic Review of Research from 2012 to 2021

Student ratings of instruction (SRI) are commonly used to evaluate courses and instructors' teaching in higher education. Students are asked to provide feedback, usually near course end dates, on their experiences in particular courses with particular instructors (Linse, 2017). SRI questionnaires typically contain a combination of questions that require students to respond to items using Likert or other rating scales and open-ended questions that allow students to articulate their perceptions and opinions in their own words. The primary purpose of SRIs is to provide instructors with formative feedback that can be used to develop teaching skills and make course improvements over time. There are instances, however, in which results have been used for hiring and promotion decisions (Becker & Watts, 1999; Centra, 1976; Medina et al., 2019). The validity and usefulness of SRIs to evaluate teaching practices for hiring and promotional purposes has been debated (Benton & Cashin, 2014; Clayson, 2009; Marsh, 2007; Spooren et al., 2013) due to concerns that SRIs may be biased by factors unrelated to teaching quality (Spooren et al., 2013). The objective of this systematic review was to identify the most recent peer-reviewed original research on potential sources of bias in SRI processes and to provide an updated and comprehensive review of bias in SRIs.

Background

Bias in SRIs has been defined as an instance "when a student, teacher, or course characteristic affects the evaluations made, either positively or negatively, but is unrelated to any criteria of good teaching, such as increased student learning" (Centra, 2003, p. 7) and is often based on uncontrollable factors such as gender, ethnicity, and physical attractiveness. For example, gender bias in SRIs has been studied extensively, and much of this research has revealed that students rate female instructors lower than male instructors (Al-Maamari, 2015; Arrona-Palacios et al., 2020; Chávez, 2020; Fan et al., 2019; Fassiotto et al., 2018; Flegl & Andrade Rosas, 2019; Martin, 2016; Mitchell & Martin, 2018; Radchenko, 2020; Wagner et al., 2016). Gender bias has been argued to reflect gender differences in teaching assignments and conditions (Arreola, 2007; Centra, 2009; Gravestock & Gregor-Greenleaf, 2008; Wright & Jenkins-Guarnieri, 2012), such as the tendency to assign more introductory course teaching assignments to women than to men (Theall & Franklin, 2001).

Racial and cultural bias may also be apparent when students rate non-white (vs white) instructors (McPherson & Jewell, 2007) and instructors with non-English (vs English) backgrounds (Fan et al., 2019) lower. SRI scores can also vary according to instructor age, teaching experience, and number of publications, where students judge younger, less experienced, and untenured instructors with fewer research publications unfairly compared to older, experienced, and tenured instructors with more publications (Clayson, 2009; McPherson & Jewell, 2007). Certain personality characteristics (Braskamp & Ory, 1994; Centra, 1993; Ferguson-Patrick, 2011), instructor likeability, class size, course level and difficulty, discipline, and delivery method (Clayson, 2009; Galbraith et al., 2012) may also influence student evaluations of their courses and instructors. Additional factors such as academic performance may influence student ratings but the positive association between these two variables may be stronger for education and humanities courses than for business and marketing courses (Clayson, 2009). These and other factors are frequently cited as reasons to avoid SRIs as a method of teaching evaluation.

Objective

The objective of our systematic review was to identify the most recent peer-reviewed original research on potential sources of bias in SRI processes, during the completion of course evaluations by students or the interpretation of course evaluations by instructors and administrators. Remaining abreast of current issues related to SRI use in higher education is important to ensure that institutional policies and procedures are aligned with evidence-informed practices for using SRIs as a method for gathering feedback on teaching and learning. In addition, as higher education policies for equity, diversity, and inclusion (EDI) continue to gain ground, it is critical to understand the role of bias in relation to SRIs. Although previous reviews have summarized the reliability, validity, stability, and biasing factors related to SRI use in higher education (Benton & Cashin, 2014; Clayson, 2009; Spooren et al., 2013; Theall & Franklin, 2001), no systematic reviews have been conducted on the topic of bias within the last 10 years. As such, we restricted our search to peer-reviewed literature published between January 1, 2012, and March 10, 2021. We identified a few reviews that included research published in the last 10 years focussing on specific disciplines (Nicolaou & Atkinson, 2019; Schiekirka & Raupach, 2015), but only one touched on bias across disciplines (Heffernan, 2021). A review by Heffernan (2021) focused on broad themes derived from thematic analysis rather than examining and addressing all the biases present in the research literature and did not report the characteristics and results of individual studies, as is the goal of the present review. Thus, the present article provides a more comprehensive review of biases identified in the literature over this period and takes a systematic approach that also considers study reporting quality.

Method

Search Strategy

Our review process was based on the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) statement (Moher et al., 2009; Shamseer et al., 2015), and Reporting Standards for Research in Psychology (Appelbaum et al., 2018). We conducted an electronic search for studies published in English using the CINAHL, Educational Resources Information Center (ERIC), ProQuest (i.e., Sociological Abstracts, ABI/INFORM, EconLit, Worldwide Political Science Abstracts), PsycINFO, Science

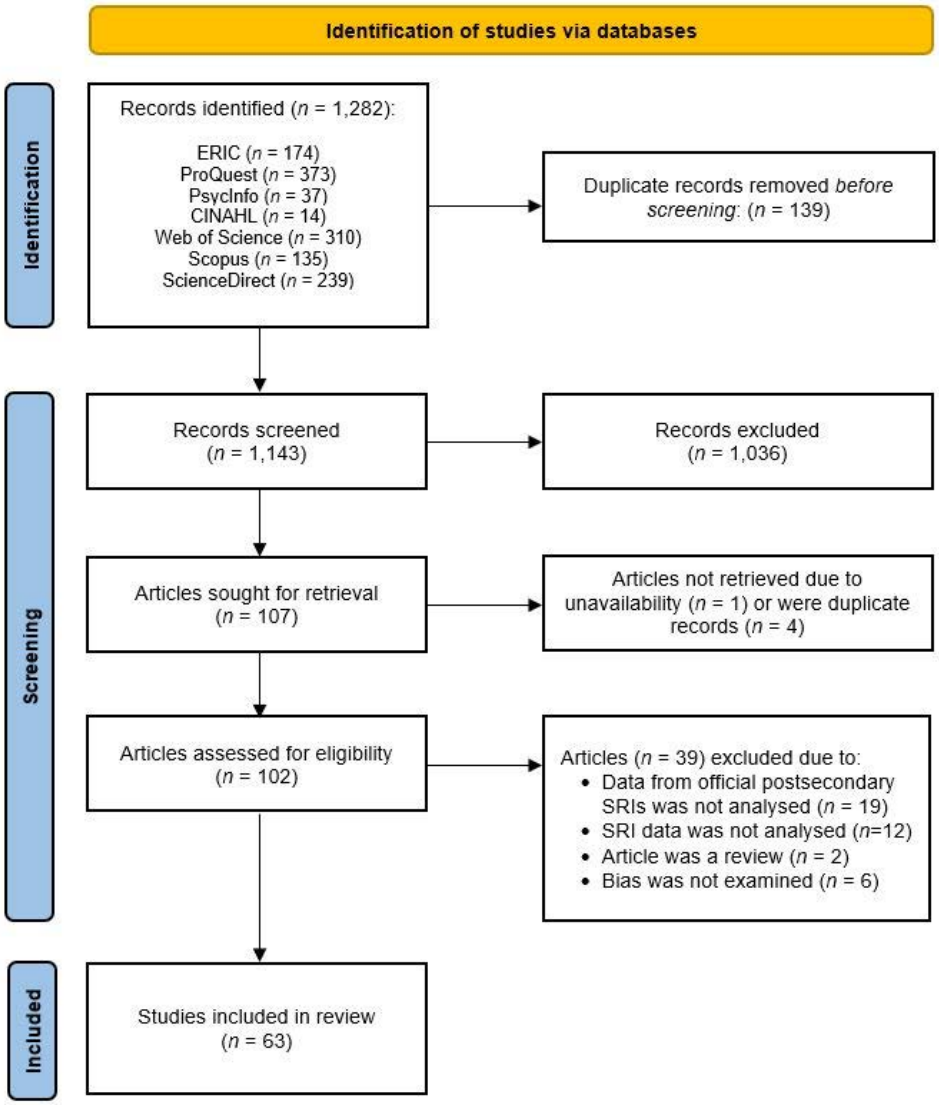
Direct, Scopus, and Web of Science databases. Our search used the following keywords and strategy: (a) (student* rating* of instruct*) OR (student* evaluation*) OR (course evaluation*) OR (student evaluation of teach*) OR (teach* effectiveness evaluation); AND (b) (biasing) OR (biased) OR (sexism) OR (prejudice) OR (discrimination) OR (implicit bias*); AND (c) (higher education) OR (postsecondary education) OR (post-secondary education) OR (tertiary education) OR (college) OR (university). The search was performed on March 10, 2021.

Selection of Studies for Targeted Review

In total, 1,282 entries were identified for the period ranging from January 1, 2012, to March 10, 2021 (see Figure 1). Research Information Systems (RIS) files, containing bibliographic citations, were downloaded from each database after each search and uploaded to Rayyan (rayyan.ai) (Ouzzani et al., 2016). Rayyan is a free web-based tool designed to help researchers working on systematic reviews and other knowledge synthesis projects and was used to screen and select studies for in-depth full text review. Using Rayyan, 139 duplicate articles were identified, reviewed, and removed.

Figure 1

Flow Diagram for Systematic Review Based on the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) Statement



Inclusion and Exclusion Criteria. Three reviewers (MQ, DB, RL) independently screened the titles and abstracts to determine inclusion eligibility based on the following criteria: The study must have (1) focused on post-secondary context; (2) aimed to determine whether student, instructor, or course factors influence evaluations of instructors or courses; and (3) analyzed data collected using the official SRIs administered by a postsecondary institution. In cases where any of these criteria were unclear or the abstract was missing, the article was included for full text review. Articles that reported the analyses of data available from public, online course/instructor evaluations (e.g., RateMyProfessors.com), or evaluations created specifically for a research study were excluded. Book reviews, case studies, commentaries, or editorials were excluded. Reviews and meta-analyses were also excluded, but relevant articles of this type were noted. Inclusion and exclusion were based on agreement by at least two reviewers (initials of study authors withheld for blind review). In cases of non-consensus, the reviewers engaged in discussion until consensus was reached. During the screening process, 107 articles met the inclusion criteria.

Full Text Review

After screening, the full text documents of 102 articles were retrieved and examined by four researchers (BMS, MQ, DB, RL). Five documents were not retrieved as they were unavailable or identified as duplicates. We excluded 39 articles during full text examination because they did not describe the analysis of data obtained through official SRI processes administered by postsecondary institutions (n = 19), SRI data were not analyzed (n = 12), the article was a review (n = 2), or the study did not examine bias in SRI (n = 6). Two articles had overlapping samples, but the measures were different (written comments vs quantitative scales) (Arrona-Palacios et al., 2020; Okoye et al., 2020). Information extracted from each article included the country of origin, study objectives, research design, participant sampling methods, student evaluator (*N*, age, gender, level of study, ethnicity) and instructor (*N*, age, gender, rank/position, ethnicity) factors, study setting (university or college), type of bias, key findings, and author conclusions and/or recommendations. Sixty-three articles met the eligibility criteria for this systematic review. We report a subset of this information in Table 1.

Table 1

Characteristics of Reviewed Studies Examining Bias in Student Ratings of Instruction (SRI)

First Author Last Name (Year); Country of study	Number of SRIs	Students			Evaluated Instructors			Type of bias: Summary of findings	Reporting Quality Assessment Domain Scores (%)			
		N	Age (years)	Gender; Ethnicity; Level	*N*	Age (years)	Gender; Ethnicity; Position		Intro	Ps	Data	Ethics
Alauddin (2014); Australia	10,223	NR	NR	undergrad, graduate	NR	NR	NR	Personality, culture: Higher SRIs associated with instructor characteristics (e.g., organization, expertise, enthusiasm, helpfulness, respectfulness) and English (vs non-English) speaking background.	75.0	33.3	33.3	0
Al-Maamari (2015); Oman	NR	2,095	NR	English language program	2,061	NR	59% F, 41% M	Gender, class size, course type, nonresponse: Female (vs male) instructors and elective (vs required) courses received lower SRIs.	75.0	66.7	77.8	0
Arnold (2019); Netherlands	NR	765	*M* = 21.5	44% F, 56% M; 79 nationalities	133	*M* = 23.5	41% F, 59% M; 26 nationalities	Gender, ethnicity/culture: Being from high (vs low) power distance countries and teaching method (individualistic vs collectivism) influence SRIs positively. No instructor gender bias effects.	75.0	33.3	55.6	0
Arrona-Palacios (2020); Mexico	NR	103,833	NR	NR	5,083	NR	NR	Gender: Students more likely to report male (vs female) professors as their best professors.	75.0	66.7	44.4	100
Bacon (2016); USA	6,754	NR	NR	NR	NR	NR	NR	Nonresponse: Low response rates advantage instructors with high SRIs and disadvantage instructors with low SRIs.	100	33.3	66.7	0
Bahous (2018); Lebanon	NR	363	NR	41% F, 59% M; 3rd year medical students	NR	NR	NR	SRI administration procedures: Compulsory SRIs did not improve reliability or influence results but were linked to increased inattentive responding rates.	50.0	66.7	66.7	100
Bianchini (2013); France	NR	1,756	NR	undergrad, graduate	NR	*M* = 45.3	NR	Age, rank, experience, satisfaction: Lower SRIs associated with older age, higher rank, fewer publications, less experience. Students satisfied with their degrees gave higher SRI ratings.	75.0	66.7	66.7	0
Blecich (2019); Croatia	NR	333	NR	NR	NR	NR	NR	Interest in course content, grades, class size: SRIs positively influenced by students' grade and (to a lesser extent) class size.	75.0	33.3	66.7	0
Boring (2017); France	NR	4,362	*M* = 18	43% M, 57% F; Year 1 undergrads	359	*M* = 34.8	67% M, 33% F; adjuncts with various professional backgrounds	Gender: Male students gave higher SRIs to male (vs female) professors, even when student grades are considered.	75.0	66.7	55.6	100
Borkan (2017); Turkey	NR	1,235	NR	NR	NR	NR	NR	Course type expected grade: Higher SRIs from students in elective (vs required) courses and with higher (vs lower) expected grades.	25.0	33.3	55.6	100
Chávez (2020); US	NR	42	NR	NR	14	NR	11 M, 3 F; 8 white, 6 non-white; instructors	Gender, ethnicity: Female instructors and instructors of colour receive lower scores than their male and white counterparts.	75.0	66.7	55.6	0

Dodeen (2013); UAE	NR	3,661	NR	undergrad, graduate	NR	NR	NR	Gender, GPA, expected grade, class size: Male (vs female) students give higher SRIs. Higher SRIs associated with higher expected grades and small class sizes. No bias related to actual GPA.	50.0	0	33.3	100
Esarey (2020); US	NR	NR	NR	NR	NR	NR	NR	Simulation to examine administrator bias: Even when evidence suggested that SRIs were reliable and correlated with teaching quality, reliance on SRIs led to misidentification of poor and good instructors.	50.0	0	44.4	100
Estelami (2015); US	NR	182	*M* = 29.4	graduate	1	NR	NR	Nonresponse, response timing: Late responders gave lower SRIs.	75.0	66.7	88.9	100
Ewing (2012); US	NR	NR	NR	NR	5,454	NR	NR	Expected grade, class size, class time, class type, class level, response rate: Higher SRIs associated with higher expected grade, small (vs large) class size, afternoon/evening (vs morning) classes, upper-level and graduate (vs lower-level) classes, higher response rate and negatively with large lecture (vs quiz, lab, small) sections.	50.0	66.7	44.4	0
Fan (2019); Australia	523,703	77,911	NR	49.7% F, 50.3% M	3,123	NR	44.2% F, 55.8% M; 38% non-English speaking	Gender, culture: SRIs biased in favor of the dominant group (males, English-speaking faculty) by local (vs international) students. Significant bias against female instructors and instructors with non-English speaking backgrounds.	50.0	33.3	66.7	100
Fassiotto (2018); US	7,888	NR	NR	NR	1,066	NR	NR	Gender, ethnicity/culture, rank, discipline: Female (vs male) medical faculty were rated lower, particularly in specialities where female faculty were underrepresented. Ethnicity bias was not observed.	25.0	33.3	55.6	0
Feistauer (2018b); Germany	517	260	NR	81% F, 19% M	26	NR	54% F, 46% M; 43% lecturers, 23% assistant professors, 34% professors	Personality, interest in course content: Instructor likability and students' interest in course content associated with positive instructor evaluations.	50.0	66.7	44.4	100
Feistauer (2018a); Germany	3,348	292	NR	77% F, 23% M; undergrad	47	NR	61.7% F, 38.3% M; 18 lecturers, 13 assistant professors or PDFs, 9 professors	Interest in course content: SRIs for lecture-based (but not seminar-based) courses biased by students' prior interest in course and clarity of course content.	75.0	66.7	55.6	0
Fischer (2019); Germany	NR	1,716	*M* = 23.2	59% F, 41% M; studying for *M* = 4.2yrs	79	NR	NR	Personality, interest in course content, expected grades: All factors positively affected SRIs.	50.0	33.3	22.2	100
Flegl (2019); Mexico	NR	10,056	NR	NR	222	*M* = 47.4	44% F, 56% M; 94% teaching contract	Gender, age, experience, course end date: Higher SRIs associated with lower instructor age, more teaching experience, higher course level, and male gender. Instructor gender effect varied by teaching experience.	75.0	33.3	44.4	0
Fletcher (2014); Canada	NR	155	NR	NR	1	NR	NR	Course type, interest in course, SRI item phrasing: Retrospective (vs prospective) phrasing of SRI items results in lower ratings of course value.	50.0	66.7	44.4	0

Fogarty (2013); US	NR	NR	NR	NR	28	NR	NR	SRI administration procedures: Lower SRIs produced with web- (vs paper-) based surveys.	50.0	66.7	66.7	0
Gith (2020); Israel	NR	36,712	NR	27,422 Jewish, 9,290 Arab	598	NR	405 Jewish, 193 Arab; instructors	Ethnicity/culture: Students rated instructors who are members of their own cultural group higher.	50.0	33.3	44.4	0
Goos (2017); European countries	NR	28,240	NR	55% F, 45% M	1,781	NR	NR	Nonresponse: SRI completion positively influenced by who completes them (i.e., those with higher grades). SRIs associated with student grades and number of evaluated courses positively, and class size negatively.	50.0	66.7	55.6	0
Griffin (2014); US	NR	2,073	NR	NR	NR	NR	NR	GPA: Lower GPA associated with higher SRIs.	25.0	66.7	66.7	0
Gupta (2018); India	NR	112,919	NR	21.9% F, 78.1% M	NR	NR	43% F, 57% M; 10% low SES	Gender, SES: Female students provided higher ratings. Male and female instructors received higher ratings in disciplines where their gender is underrepresented and from students of the same SES.	50.0	33.3	55.6	0
Jobu Babin (2020); US	NR	2,968	*M = 23.4*	45% F, 55% M; 12.24% African, 8.16% Asian, 9.2% Hispanic; undergrad	284	*M* = 44.41	38% F; 23.2% tenured faculty	Gender, attractiveness, expected grade, teaching mode: Attractive faculty receive higher SRIs. Attractiveness effect evident in face-to-face (but not online) courses and was stronger for female (vs male) instructors. SRIs higher in face-to-face (vs online) courses and small (vs large) classes. Students taking major courses and those with higher expected grades provided higher SRIs.	50.0	66.7	77.8	100
Laupper (2020); Switzerland	NR	463	*M* = 46.7	28.1% F, 71.9% M; vocational training	NR	NR	NR	SRI administration procedures: Online SRI administration matches the data quality of paper-based delivery.	50.0	66.7	44.4	100
Liu (2012); US	NR	11,351	16-22 (16.3%), 23-30 (20.8%), 31-40 (32.5%), >40 (30.5%)	24.7% F, 75.3% M; 7.5% 1st yr, 6.5% 2nd yr, 17.3% 3rd yr, 16.9% 4th yr, 30.7% graduate	1,522	NR	62% F, 38% M; 51.7% instructors, 19.1% assistant professors, 15.8% associate professors, 13.5% professors	Level of study, course type, rank, gender, class size: First year (vs higher level) students provide lowest SRIs. Elective (vs required) courses rated higher. Assistant professors and professors rated lower than instructors and associate professors. No evidence of instructor gender, class size, part- or full-time faculty biases.	25.0	66.7	55.6	0
Macfayden (2016); Canada	NR	21,534	NR	NR	NR	NR	NR	Academic performance, gender, discipline, class size, course level: Male (vs female) students, students with higher grades, and in first year completed more SRIs. Completion rate dropped with each year. Class size negatively associated with SRI completion.	50.0	66.7	77.8	100
Magel (2017); US	NR	NR	NR	NR	387	NR	48.6% F; 98 full, 103 associate, 179 assistant professors, 9 instructors	Gender: Male salaries increase as SRIs increase. Female salaries decrease as SRIs increase.	50.0	66.7	44.4	100

Maricic (2019); Croatia	1,636	NR	NR	NR	6	NR	50% M, 50% F	Gender: Students rate different aspects of teaching as important for male (clarity, professionalism, objectivity) and female (nurturing) instructors.	75.0	66.7	44.4	0
Martin (2016); US	309	NR	NR	NR	NR	NR	NR	Gender, class size: Females typically teach smaller courses than males. Male (vs female) instructors receive higher SRIs in larger courses.	75.0	33.3	44.4	0
Mitchell (2018); US	1,424	NR	NR	NR	2	NR	1 M, 1 F; professor	Gender: SRI comments focused on women's appearance and personality and suggested lower professional respect.	75.0	66.7	55.6	0
Nargundkar (2014); US	NR	105,974	NR	undergrad, graduate	NR	NR	Tenured, nontenure track, part- time instructor, teaching assistant	Semester, time of day, course type/level, class size, instructor rank, gender: Ratings for spring/summer (vs fall), evening (vs morning, afternoon), non-core (vs core), graduate (vs undergrad), and small (vs large) classes rated higher. Undergrads rated female (vs male) instructors of non-core courses higher. Graduate students rated male (vs female) instructors higher. Higher SRI observed for non-tenure track faculty in graduate courses, and part-time instructors teaching undergrad classes.	75.0	33.3	33.3	0
Okoye (2020); Mexico	82,144	NR	NR	undergrad	NR	NR	instructor	Gender: Written comments reveal that students value male instructors for demonstrating knowledge and female instructors for teaching methodology and clear explanations.	75.0	66.7	66.7	100
Palali (2018); Netherlands	28,243	9,000	*M* = 22.0	44% F, 56% M; 48% undergrad, 52% graduate	83	*M* = 44-46	8-16% F, which varied by course	Number of publications: Undergrad (but not graduate) students gave lower SRIs to instructors with more research publications.	25.0	66.7	100	0
Park (2020); US	NR	NR	NR	17% first-year, 16% sophomore, 24% junior, 23% senior, 16% graduate	2,870	NR	48% F, 52% M; 68% white; 27% Tenure/ tenure-track, 42% adjunct, 15% term, 15% grad assistant	Gender, ethnicity/culture, faculty type, course type (elective vs required), SRI administration procedures: Making the course intellectually stimulating was more important for female than male students in predicting course effectiveness. No evidence of gender or ethnicity biases.	50.0	66.7	77.8	100
Park, H -S (2018); Korea	NR	1,206	NR	42% F, 58% M; 21% Year 1, 79% upper level	NR	NR	NR	Monotonic responding: First-year students with low course grades engaged in straight-line responding.	100	33.3	44.4	100
Peterson (2019); US	NR	NR	NR	NR	4	NR	50% M, 50% F; instructor	Gender: SRIs for female (but not male) instructors were higher when male (but not female) students were informed about potential gender bias before completing SRIs.	75.0	33.3	66.7	100
Punyanunt-Carter (2015); US	NR	58	NR	48.3% F, 51.7% M	NR	NR	NR	Gender: Male students rate female (vs male) instructors higher. Female students rate male (vs female) instructors higher.	75.0	66.7	11.1	0

Radchenko (2020); US	NR	365,187	NR	61% F, 39% M; undergrad, graduate	2,093	NR	40% F, 60% M; 60% full time professors	Gender, course type, class size: SRIs are lower for female (vs male) instructors, graduate (vs undergrad) courses, required (vs elective) courses, medium and large (vs small) class size, lower (vs higher) expected grade. Match between student and instructor gender raises SRIs.	75.0	66.7	77.8	100
Reisenwitz (2016); US	NR	313	$M = 20.6$	42.5% F, 57.5% M; 60% White; undergrad	NR	NR	NR	Gender, time poverty, complaining behavior, academic performance, ethnicity/culture, technology savviness: Male students and those with higher grades are more likely to complete SRIs. Time poverty, complaining behavior, and technology savviness did not influence SRI completion.	75.0	33.3	77.8	100
Risquez (2015); Ireland	63,173	NR	NR	NR	673	NR	NR	SRI administration procedures, class size, preparation: Delivery mode (paper vs online) had no effect on SRIs after controlling for class size, faculty.	50.0	66.7	55.6	0
Rodríguez (2014); Spain	NR	1,359	$M = 20.3$	57.1% men, 42.9% women	125	NR	NR	Instructor age, gender, experience, grades, class size: Teacher experience and pedagogy positively related to students' perceptions. Students perceive men to have more expertise, but women to have better attitudes. Younger instructors are perceived to have better attitudes. Grades positively and class size negatively correlated with student perceptions.	100	33.3	55.6	0
Royal (2015); US	NR	2,564	NR	NR	NR	NR	NR	Course type: Students were more critical of instructors of methods (vs non-methods) courses.	50.0	33.3	55.6	0
Schönrock-Adema (2013); Canada, Netherlands	NR	966	NR	NR	NR	NR	NR	Grade expectation: Students expecting high (vs low) exam scores were more satisfied. Female (vs male) students and students who were more satisfied after an exam rated courses more highly.	50.0	66.7	66.7	100
Schueths (2013); US	NR	NR	NR	NR	29	$M_F = 42.6$, $M_M = 46.6$	20.6% ethnic minority, 79.3% white; 6 teaching assistants, 3 lecturers, 10 professors, 4 other	Diversity: Minority, particularly female instructors of colour, evaluated as more biased than their white and male instructors teaching similar curricula. Non-minority instructors may gain privileges in the evaluation process by avoiding topics in diversity.	50.0	66.7	33.3	0
Socha (2013); US	NR	4,063	NR	undergrad, graduate	89	NR	NR	Course difficulty, interest in course content, expected grade, ethnicity, age, level of study: Higher prior interest associated with lower SRI. Higher grade expectations associated with higher SRIs. White (vs non-white) instructors and young (vs old) instructors received higher ratings. Undergrad (vs graduate) courses and those with higher (vs lower) workload received lower SRIs. Pace of the course, student gender, reason for taking the course, course credit hours, course enrolment, course average grade, teacher gender and teacher rank did not influence SRIs.	25.0	66.7	66.7	0

Spooren (2017); Belgium	NR	927	NR	66% F, 34% M; undergrad	NR	NR	NR	Gender, personality, discipline: SRI completion not influenced by instructor likeability, teaching skills, or personality, course workload. Male (vs female), first-year (vs sophomores), and natural and life science students value SRIs more, leading to higher SRIs.	100	66.7	55.6	0
Sulis (2019); Italy	6,425	NR	NR	65.5% F, 34.5% M	55	NR	24.9% Full, 39.5% associate, 35.6% assistant	Interest in course content, instructor rank: SRIs influenced positively by students' prior interest and knowledge of course content. SRIs not associated with instructor rank.	50.0	66.7	33.3	0
Tarun (2016); US	209	209	NR	undergrad	3	NR	NR	Test type, grades, interest in course content, course difficulty/workload: SRIs influenced by assessment type, grades, interest in course, course difficulty, and student workload.	50.0	33.3	22.2	0
Tomes (2019); South Africa	NR	257	NR	63% F, 37% M; African, mixed-race, Indian, Asian, White; undergrad	1	NR	NR	Gender, inattentive responses, nonresponse, ethnicity/culture, academic performance: Male (vs female) students and those with higher academic performance gave higher SRIs.	75.0	33.3	33.3	0
Treischl (2017); Germany	NR	2,037	NR	NR	NR	NR	NR	SRI administration procedures: Marked differences in non-response rate between online and paper-based SRI, small differences in SRIs with online version having more negative ratings.	25.0	66.7	44.4	0
Valencia (2020); Canada	NR	3,000	NR	graduate	NR	NR	71% M, 29% F	Gender: Female (vs male) instructors received higher SRIs when accounting for acquiescence in student responses.	75.0	66.7	66.7	100
Wagner (2016); Netherlands	NR	NR	NR	NR	93	*M* = 48	43% F, 57% M; 33% non-Caucasian	Gender, ethnicity: Negative female instructor effect on SRIs even when controlling for instructor characteristics. No ethnicity effect.	50.0	33.3	66.7	0
Wang (2020); US	915	NR	NR	24.6% F, 75.4% M; 45.5% White, 48.5% foreign origin	264	NR	128 foreign origin; full-time professors	Gender, ethnicity/national origin, course difficulty: SRIs biased against instructors with minority racial, ethnic, and foreign cultural backgrounds. No gender effects.	75.0	66.7	44.4	0
Weidman-Evans (2020); US	NR	265	NR	NR	NR	NR	NR	Grades, course difficulty/workload: SRIs were not related to grades or credit hours.	25.0	33.3	55.6	0
Winer (2016); Canada	NR	18,669	NR	Responders: 59% F, 41% M; non-responders: 41-43% F, 59% M; undergrad	NR	NR	NR	Nonresponse, discipline, class size, timing of SRI, feelings about the course, rigorous grading: No evidence to support concerns about the validity and usefulness of online SRIs. Academically stronger students responded at a higher rate and smaller courses received more favourable SRIs.	100	66.7	66.7	0
Wolbring (2012); Germany	18,000	NR	NR	NR	NR	NR	NR	Absenteeism: Number of classes missed decreases with increasing teaching quality. When adjusting for bias due to absenteeism in course rankings based on SRI, average courses are more strongly affected than courses of very high or low quality.	75.0	33.3	77.8	0

Wolbring (2016); Germany	NR	1,335	NR	NR	NR	NR	NR	Nonresponse: Positive climate among students reduces absenteeism whereas increased course load and workload increases absenteeism. Greater absenteeism associated with slight increase SRI scores. SRI-based ranking of courses was affected dramatically.	50.0	0	44.4	100
Yueh (2012); Taiwan	NR	3,125	NR	10.6% F; 18% Year 1/2, 26.5% Year 3, 55.5% seniors & graduate	NR	NR	NR	Expected grade, attendance, level of study, institution type: Higher (vs lower) attendance, higher (vs lower) grade expectations, seniors and graduate (vs Years 1, 2, 3), and those in universities (vs. public technology colleges) provided higher SRIs.	50.0	33.3	22.2	0

Note. M = male, F = female, Intro = Introduction and Ps = Participant domains of the Quality of Survey Studies in Psychology (Q-SSP) Checklist (Protogerou & Hagger, 2020), PDFs = post-doctoral fellows, SES = socio-economic status

Quality of Reporting Assessment

The quality of reporting is viewed as essential when synthesizing research evidence. Therefore, we examined the quality of reporting using the 20-item Quality of Survey Studies in Psychology (Q-SSP) Checklist, which was developed and validated through a multistep procedure using an expert-consensus method (Protogerou & Hagger, 2020). The Q-SSP assesses reporting practices in four domains: Introduction (rationale, variables; 4 items), Participants (sampling; 3 items), Data (collection, analyses, measures, results, discussion; 10 items), and Ethics (3 items) within an article (Protogerou & Hagger, 2020). Items are rated as 1 = *Yes*, 0 = *No* or *Not Stated Clearly*, or *Not Applicable*. *Domain* and *Overall Quality* of reporting scores in the form of percentages are calculated from the number of "Yes" codes divided by the number of applicable items. Reporting quality is considered acceptable when the percentage is equal or greater than 70% (depending on the number of applicable items). Article reporting quality was examined by at least two researchers. Individual ratings were compared on an item-by-item basis, and percent agreement was computed (M = 86.8%, SD = 7.7%, *Range* = 76.2 – 100%). In cases of non-consensus, the reviewers engaged in discussion but if this did not result in consensus, a third reviewer joined discussions and provided tie-breaking ratings.

Results

Study Characteristics

Studies were conducted in the USA (n = 25), Germany (n = 6), Canada (n = 4), Netherlands (n = 4), Canada and the Netherlands (n = 1), Mexico (n = 3), Australia (n = 2), Croatia (n = 2), France (n = 2), and various other regions (n = 15). Most studies conducted quantitative analyses (n = 61) based primarily on cross-sectional surveys administered at multiple time points. Two articles described qualitative approaches to examining responses to open-ended questions in SRIs. Thirty-eight articles reported that data obtained from 261,507 students were analyzed. Twenty-five articles reported the number of SRIs analyzed (N = 1,495,957). Eight of these studies reported both student sample sizes and the number of SRIs analyzed. Neither student sample size nor number of SRIs were reported in eleven articles but seven of these studies reported the number of instructors, five reported the number of courses (N = 7,116) and relied on course-level data, and one was a simulation study drawing from an existing database. In total, 42 articles reported information about 28,659 instructors.

Quality of Reporting Assessment

The overall reporting quality of the reviewed studies was relatively low (M = 54.3%, SD = 12.3%, *Range* = 29.4 – 82.4%). These overall scores obscured ratings of acceptable reporting practices in at least one domain of 44 studies. We summarized the strengths and weakness of the studies in each domain to capture quality reporting (see Table 1). Acceptable quality was observed in the *Introduction* domain for 29 studies (46%) (M = 60.3%, SD = 20.4%, *Range* = 25.0 – 100%). All articles described and justified the problem under investigation but many (n = 35, 55.6%) overlooked describing the relevant population. This finding, combined with lack of explanation of the research questions and variables examined (n = 31, 49.2% and n = 34, 54.0%, respectively), contributed to reduced scores in the Introduction domain. None of the articles showed acceptable reporting quality in the *Participant* domain (M = 50.8%, SD = 19.7%, *Range* = 0 – 66.7%), which was largely due to the lack of sample size justification. However, most studies stated participant inclusion criteria (n = 48, 76.2%) and recruitment strategies (n = 47, 74.6%). Acceptable reporting quality was observed in 9 (14.3%) articles in the *Data* domain (M = 54.7%, SD = 17.2%, *Range* = 11.1 – 100%).

Articles with higher quality scores adequately justified the analytic techniques (n = 54, 85.7%), provided information about study context (n = 47, 74.6%), and described the findings in relation to the appropriate population (n = 62, 98.4%). Articles with reduced reporting quality scores overlooked reporting of the completion rate (n = 51, 80.0%), treatment of missing values (n = 54, 85.7%), or provided few demographic details about the study sample (n = 58, 92.1%). For the *Ethics* domain, 23 (36.5% of articles) showed acceptable quality in reporting (M = 36.5%, SD = 48.5%, *Range* = 0 – 100%). However, two items were not considered relevant for studies involving secondary data analysis, which included nearly all studies in our review. This reduced the number of applicable items in the *Ethics* domain to one.

Types of Bias

Studies examined biases associated with gender (*n* = 31, 49.2%), class size (*n* = 13, 20.6%), ethnicity/culture (*n* = 11, 17.5%), nonresponse (*n* = 10, 15.9%), expected grades (*n* = 9, 14.3%), interest in course content (*n* = 7, 11.1%), SRI administration procedures (*n* = 6, 9.5%), academic performance (*n* = 6, 9.5%), and course type (elective vs required) (*n* = 6, 9.5%). Other types of bias included class level, discipline, course difficulty/workload, instructor rank, age. Thirty-eight articles (60.3%) examined more than one type of bias.

Gender. A few studies did not find evidence of gender bias. Flegl and Andrade Rosas (2019) found gender differences in SRIs were no longer evident when they controlled for instructor age and experience. Liu (2012) reported that neither instructor nor student gender predicted SRI scores. Using decision tree analysis, E. Park and Dooris (2020) found that instructor gender did not predict SRI scores. Despite these null findings, most reviewed studies provided evidence of gender bias in SRIs (*n* = 27, 42.9%), with either student gender (*n* = 6), instructor gender (*n* = 19), or an interaction between the two (*n* = 2) having an influence on SRIs. Most articles examining instructor gender indicated bias against female instructors (*n* = 16). Students were more likely to recommend males over females when ranking an individual as their best professor (Arrona-Palacios et al., 2020) and provided lower SRI scores to female than male instructor and professors (Al-Maamari, 2015; Chávez, 2020; Fassiotto et al., 2018; Radchenko, 2020; Wagner et al., 2016), despite similar average exam scores in courses taught by female and male professors (Boring, 2017). Interestingly, when gender bias or acquiescence were considered, SRI scores for female instructors increased (Peterson et al., 2019; Valencia, 2020). One study found evidence of same-gender bias in ratings, with male students rating male instructors higher than female instructors (Boring, 2017). Evidence of cross-gender bias was also noted whereby male students rated female instructors higher than male instructors and vice-versa (Punyanunt-Carter & Carter, 2015). SRI scores were also biased towards female instructors' attractiveness in face-to-face courses compared to online courses (Jobu Babin et al., 2020). When potential bias was made salient through anti-bias instructions, male (but not female) students' ratings of female (but not male) instructors increased (Peterson et al., 2019).

Five studies (7.9%) examined student comments in SRIs, which revealed a stark contrast between genders. Mitchell and Martin (2018) examined over 82,000 comments made in SRIs. Comments focused on female instructors' appearance and personality and demonstrated lower levels of professional respect compared to comments made about male instructors. Maricic et al. (2019) found that student ratings of an instructors' clarity, professionalism, and objectivity were more important for male professors in predicting overall impressions, whereas nurturing qualities were emphasised for female professors. Moreover, male instructors were valued for demonstrating knowledge and expertise and female instructors for teaching methodology, providing clear explanations, and attitude (Okoye et al., 2020; Rodríguez et al., 2014). Downstream effects of bias related to SRIs were demonstrated in decisions related to salary increases, with male but not female salaries increasing based on higher SRI scores (Magel et al., 2017).

Ethnicity and Culture. Eleven (17.5%) studies reported evidence of bias against faculty with minority ethnic and cultural backgrounds. Instructors of colour received lower SRI scores than their White counterparts (Chávez, 2020; Socha, 2013; Wang & Gonzalez, 2020). Further, minority instructors, particularly female instructors of colour, who taught required diversity courses were judged more negatively than were non-minority male and female instructors by students, whose comments focused on the minority instructors' bias (Schueths et al., 2013). Students' own cultural background also influenced their ratings (Arnold & Versluis, 2019) and were biased in favor of instructors with similar backgrounds (Gith, 2020). This bias may be pronounced when there is less diversity within a faculty (Fan et al., 2019). Finally, English-speaking (vs non-English-speaking) faculty were rated more highly, especially by domestic (vs international) students (Fan et al., 2019). In contrast, there was little evidence of an ethnicity bias in SRI scores when an institution was considered highly diverse, brought together instructors and students from across the globe, and focused on social justice as part of their mission (Wagner et al., 2016)(Wagner et al., 2016).

Other Instructor Characteristics. First impression of instructors, and instructor characteristics such as enthusiasm, organization, interesting presentation style, providing adequate feedback, content expertise, providing clear explanations, treating students with respect, and humor were positively associated with SRI scores (Alauddin & Kifle, 2014; Fischer & Hänze, 2019). Students who were fond of

their instructors provided higher SRI scores than those not fond of their instructors (Feistauer & Richter, 2018b); this association was not related to SRI completion rate (Macfadyen et al., 2016). Students also provided higher ratings to instructors with more years of experience, but the benefit of experience was limited as older professors were rated lower (Bianchini et al., 2013; Flegl & Andrade Rosas, 2019). In contrast, one study showed little evidence of an association between instructor rank and SRI scores (Sulis et al., 2019).

Student Factors. Seven (11.1%) studies examined the influence of students' prior interest in course content on their SRI scores. Prior interest was positively related with SRI scores (Blecich & Zaninović, 2019; Feistauer & Richter, 2018a, 2018b; Fischer & Hänze, 2019; Sulis et al., 2019; Tarun & Krueger, 2016) and students with a greater prior understanding of the course content provided higher ratings (Sulis et al., 2019). Conversely, Socha (2013) found that prior interest level was negatively associated with SRI ratings, possibly mediated by overly high course expectations. Further, students provided higher ratings on course evaluations in elective (vs required) courses, which could also be attributed to their understanding and prior interest in the course (Al-Maamari, 2015; Borkan, 2017; Liu, 2012; E. Park & Dooris, 2020; Radchenko, 2020). Courses perceived as more difficult were rated lower (Tarun & Krueger, 2016) and instructors of quantitative methods courses, which are often perceived as difficult, were rated lower than instructors of other courses, even though many aspects of quantitative methods courses were preferred by students (Royal & Stockdale, 2015). Students also evaluated higher-level courses more positively than lower-level courses (Ewing, 2012; Flegl & Andrade Rosas, 2019; Liu, 2012; Nargundkar & Shrikhande, 2014; Socha, 2013; Yueh et al., 2012).

Fifteen (23.8%) studies examined the effect of student grades and academic performance on SRIs. Students with academically strong backgrounds or higher GPAs were more likely to complete SRIs (Macfadyen et al., 2016; Reisenwitz, 2016) and rate instructors more positively (Fischer & Hänze, 2019; Rodríguez et al., 2014; Tarun & Krueger, 2016; Tomes et al., 2019; Winer et al., 2016). Further, students expecting higher grades provided higher SRI ratings (Blecich & Zaninović, 2019; Borkan, 2017; Dodeen, 2013; Ewing, 2012; Goos & Salomons, 2017; Jobu Babin et al., 2020; Radchenko, 2020; Socha, 2013; Tarun & Krueger, 2016; Yueh et al., 2012). The influence of expected grades on SRI is likely to interact with other biases, in particular non-response bias, given that non-responders are more likely to have lower grades (Reisenwitz, 2016). In two studies, associations between actual grades (Weidman-Evans et al., 2020) and GPA (Dodeen, 2013) with SRI scores were not evident.

First year students were most likely to complete SRIs, but response rates declined as they advanced through their tenure, suggesting an "evaluation fatigue" effect (Macfadyen et al., 2016; Spooren & Christiaens, 2017). SRIs are also influenced by the quality of responses to SRI items. H.-S. Park and Cheong (2018) described the prevalence of monotonic (straight line) response patterns among first-year students with lower grades. This response pattern was attributed to "lower level of motivation, lack of familiarity with the course evaluation process, and/or inadequate understanding of the importance of course assessment to university decisions" (H.-S. Park & Cheong, 2018, p. 109).

Students from high (vs low) power distance countries gave higher ratings to instructors (Arnold & Versluis, 2019). Power distance refers to "the extent to which the less powerful persons in a society accept inequality in power and consider it as normal" (Hofstede, 1986, p. 307). Further, students favoured instructors with similar socio-economic backgrounds, rating them higher (Gupta et al., 2018). Pedagogical methods and the cultural background of students may also interact, resulting in skewed SRI scores. For example, pedagogies consistent with individualistic cultural lens were rated higher by students from individualistic cultures as opposed to those from collectivist cultures (Arnold & Versluis, 2019). Lastly, SRI scores were not influenced by students with a greater sense of time poverty, predisposed to complain, or technological savviness (Reisenwitz, 2016).

Course Factors. The effect of class size on SRIs was observed in 13 (20.6%) studies. In most studies, class size negatively influenced SRIs, such that smaller courses were rated more favourably (Blecich & Zaninović, 2019; Dodeen, 2013; Ewing, 2012; Goos & Salomons, 2017; Jobu Babin et al., 2020; Macfadyen et al., 2016; Nargundkar & Shrikhande, 2014; Risquez et al., 2015; Rodríguez et al., 2014; Winer et al., 2016). Gender biases in SRIs also increased with class size, with bias against female instructors being most evident in large classes (Martin, 2016). Liu (2012) and Bianchini et al. (2013) did not find a significant impact of class size on SRI scores.

SRI Administration Procedures. Wolbring and Treischl (2016) found that the timing of the SRI

(first or last of the day of the course) led to low SRI response rates, which was partly attributed to dissatisfied students not attending class to contribute to the overall rating of courses. Moreover, students who responded later (vs earlier) in a course evaluation period provided lower ratings (Estelami, 2015). This latter finding suggests the possibility of the introduction of a positive bias if late responders were unable to participate in the SRI process at all (Bacon et al., 2016). To tackle non-response bias, one study examined the effect of mandatory SRI completion; this action was associated with reduced reliability of SRI scores and increased rates of inattentive responding (Bahous et al., 2018). SRI scores may also be skewed by the timing of the course within a day or academic year. Nargundkar et al. (2014) found that ratings were significantly higher for summer and spring (vs fall) semesters and instructors were rated higher for evening (vs morning and afternoon) classes. Ewing (2012) and Wolbring (2012) found that morning classes received significantly lower ratings compared to evening and afternoon classes.

Little or no influence of SRI delivery mode (paper-based vs electronic/online) on overall SRI scores was found in three studies (Laupper et al., 2020; Risquez et al., 2015; Treischl & Wolbring, 2017). In contrast, Fogarty et al. (2013) found significantly lower evaluation scores with web-based SRI administration. Lower ratings in evaluations, however, may be due to reduced response rates in asynchronous online vs paper based administered SRIs (Treischl & Wolbring, 2017). When instructors provided students with class time to complete SRIs, however, some boost in response rates were observed (Risquez et al., 2015; Treischl & Wolbring, 2017).

Discussion

The results of this systematic review indicate that bias can be introduced into SRIs by factors that are unrelated to the course, or the quality of teaching and our overall findings provide additional support for themes identified in a recent literature review (see Heffernan, 2021). The existence of gender bias was the most consistent and prominent finding across the studies we reviewed. Student ratings of female instructors were lower than those for their male counterparts (Al-Maamari, 2015; Arrona-Palacios et al., 2020; Chávez, 2020; Fan et al., 2019; Fassiotto et al., 2018; Flegl & Andrade Rosas, 2019; Martin, 2016; Mitchell & Martin, 2018; Radchenko, 2020; Wagner et al., 2016) and written comments for female instructors used less professional language and were more likely to focus on appearance and personality (Mitchell & Martin, 2018). Interestingly, when students are made aware of issues related to gender bias, male students rated female instructors higher than when they were not made aware of such bias (Peterson et al., 2019). Bias against instructors with minority racial, ethnic, and foreign cultural backgrounds was another substantive finding in our review (Chávez, 2020; Gith, 2020; Schueths et al., 2013; Socha, 2013; Wang & Gonzalez, 2020), particularly in institutions with less diversity among faculty and students (Fan et al., 2019). SRIs were also positively related to factors such as students' prior interest in course content (Blecich & Zaninović, 2019; Feistauer & Richter, 2018a, 2018b; Fischer & Hänze, 2019; Sulis et al., 2019; Tarun & Krueger, 2016) and grade expectations (Borkan, 2017; Goos & Salomons, 2017; Radchenko, 2020), and negatively associated with class size (Blecich & Zaninović, 2019; Goos & Salomons, 2017; Jobu Babin et al., 2020; Macfadyen et al., 2016; Risquez et al., 2015; Winer et al., 2016).

Implications

The existence of bias in SRIs should not be ignored as inaccurate SRI results can have serious implications for instructors, especially for those who find themselves at the intersection of multiple biases (e.g., females who come from marginalized or minority groups). Incorporating evidence-informed practices when using SRIs to evaluate courses or instructors is extremely important.

Many researchers argue that the gender bias is a critical concern (Boring, 2017; Fan et al., 2019; Fassiotto et al., 2018; Mitchell & Martin, 2018; Radchenko, 2020). Others have concluded that gender bias is unlikely to have a substantial impact on SRIs given the small effect sizes found in a few studies (Al-Maamari, 2015; Arrona-Palacios et al., 2020; Benton & Cashin, 2014). Even small effect sizes, however, can have a meaningful impact especially when high stakes outcomes are involved (e.g., hiring, salary, tenure, and promotion decisions) and the bias can systematically disadvantage some faculty, particularly women and instructors of colour, relative to others. Bias is problematic when instructors' performance is compared against an arbitrary criterion that demarcates acceptable from unacceptable performance (Wagner et al., 2016). Use of inaccurate SRI results for making personnel decisions may also foster a cul-

ture of manipulation of the ratings by instructors through means of easy tests, lenient grading, and other incentives for students (Stroebe, 2016; Wolbring & Treischl, 2016). This reward/punishment system is evident in the relationship between students' expected grades and their SRIs, which may not be reflective of actual teaching abilities at all (Borkan, 2017; Radchenko, 2020). These results support theories that suggest that, when used for career progression, SRIs are a contributing factor to the underrepresentation of women in fully tenured and university leadership positions (Fan et al., 2019; Heffernan, 2021) and the reduced retention of faculty members from diverse backgrounds (Boring, 2017). More research is needed to understand how SRIs are interpreted and used by institutional decision makers and the impact on specific groups of instructors.

Bias in student evaluations can result in cumulative bias in the interpretation of SRI results. SRIs do not work well when comparing instructors to one another, or when examining only one set of results at a single timepoint. Instructors and administrators are encouraged to "avoid appraisals on [SRI] based on the observation of a single academic year for lecturers who have been teaching in the same institution for more academic years" (Sulis et al., 2019, p. 1328). In addition, means or medians tend to be the metric of choice when interpreting SRI results, but the use of these norm referenced measures often means that the distribution of ratings across SRI scales are ignored (Linse, 2017). Ratings across the spectrum, however, may provide valuable information about strengths and areas requiring further development (Medina et al., 2019). Benchmarking and comparisons across disciplines and courses is also unwise as student ratings can vary widely in these domains (Benton & Ryalls, 2016). Comparisons using SRIs are most useful and appropriate when they are made within the set of ratings for a single instructor and observing how these ratings change over time (Medina et al., 2019).

Comparing SRIs across courses that differ in class size, course-level, required status, difficulty, and students' prior interest is also problematic. Instructors teaching larger introductory courses may be disadvantaged, as smaller courses are (in some ways) easier to teach (e.g., reduced grading demands, fewer students to interact with and assist). Likewise, instructors teaching more difficult courses and courses where students have lower prior interest may also be disadvantaged. Benton and Cashin (2014) argued that these factors should be statistically controlled or that groups are matched on these characteristics for comparison. These options, however, are difficult to implement practically considering the significant number of biases that need to be taken into account if these approaches are used (Royal & Stockdale, 2015).

A critically important consideration is the use of SRIs to measure teaching effectiveness. A simulation study highlighted that even when SRI scores appear valid and reliable, they can often mis-identify poor and good instructors (Esarey & Valdes, 2020). We echo other researchers in arguing that it is imperative that instructors use more than this lone source of information to inform their teaching development plans (Flegl & Andrade Rosas, 2019; Weidman-Evans et al., 2020). Bias in SRIs also effects the usefulness of SRI results to inform and enhance instructors' teaching practices and make course improvements, which may impact students' educational experiences and outcomes. This is problematic for women students and students of colour and foreign cultural background, whose own academic motivation and outcomes are improved when taught by an instructor who shares their identity (Carrell et al., 2010; Fairlie et al., 2014; Hoffman & Oreopoulos, 2009; Llamas et al., 2021).

Considering the nature of the questions when using SRIs as a measure of teaching effectiveness is important. Ray et al. (2018) examined 1,074 questions in 55 SRIs from 270 postsecondary institutions and found that instructors were the subjects of many questions. Focussing on instructors rather than the course components or design opens the door for students to make biased judgements based on the uncontrollable instructor characteristics we have discussed. Recent studies have recommended that SRI questions be phrased to focus on student learning and engagement rather than instructor performance. Newer SRI questions shift the focus towards students' experiences in areas of teaching and learning that they were more likely to evaluate appropriately, thereby minimizing the effect of bias (Centre for Teaching Support & Innovation, 2018). Preliminary studies suggest that systematic gender, faculty rank, age, or seniority biases have been reduced (Centre for Teaching Support & Innovation, 2018) and questions asking students to evaluate the course rather than the instructor may reduce gender and cultural biases (Fan et al., 2019) as well. Further reductions in bias may be accomplished by increasing students' awareness of biases that exist in instructor evaluations (Fan et al., 2019; Peterson et al., 2019). Educating staff and administrators about the need to address various sources of bias in policies and to be thoughtful

during tenure and promotion decision-making processes if considering SRIs is also important (Magel et al., 2017). Future studies should examine the ways in which SRI results are applied across institutions and explore ways to reduce bias when interpreting SRI reports.

Strengths and Limitations

The findings from our review extended and supported the work of previous systematic reviews on this topic. The current review was limited to peer-reviewed, original research studies published in the last 10 years; however, we did not ignore the earlier literature and consulted reviews (Spooren et al., 2013) to determine whether similar themes were present in the past research. Indeed, the types of bias that we observed in the reviewed studies have been described previously, suggesting that advances in education, EDI initiatives, and technology have not had a major impact on the presence of bias in student evaluations of their instructors and courses. One limitation of our systematic review is that we did not consult the grey literature, which may have included studies reporting null findings (Rosenthal, 1979). One might argue that this decision biased our own review towards significant effects reported in the peer-reviewed literature. We suspect that should this bias exist, it is minimal given the large sample sizes of the reviewed studies and that null results were present.

The attention given to issues related to EDI is a strength of our review. We found two areas, in particular, that require further attention in research. First, we found no reference to individuals who do not identify as men or women and discussions of bias as it relates to non-heteronormative instructors or students does not appear in the relevant SRI literature. This is a limitation of most studies related to bias in SRI. Definitions of gender as a binary variable synonymous with biological sex (male or female) need to be reconceptualized and research in this area should include definitions of gender identity and gender expression (see Lindqvist et al., 2020) to examine the potential for bias directed toward instructors with non-binary identities. Second, most of the reviewed studies showed evidence of ethnicity bias in SRIs, but the findings of Wagner et al. (2016) diverged from this pattern. Lack of an ethnicity bias in SRI scores were attributed, in part, to the Dutch institution's faculty and student global recruitment efforts and focus on social justice(Wagner et al., 2016). Given the increasing emphasis on EDI in higher education around the world, future research should examine the impact of these initiatives (including the increased adoption of non-White/non-Western pedagogies, and changes in recruitment and hiring practices) on students' evaluations of teaching, especially for groups of instructors (e.g., women, instructors of colour) who have received biased SRI reports. Research is also required to examine the typical approach to gathering student feedback on teaching quality and their learning experiences, and value other ways of knowing and defining "what will count" as quality instruction (Louie et al., 2017).

Another strength of our review is that we examined the level of quality in reporting across the studies that we reviewed. The studies that we examined demonstrated reasonable levels of reporting quality within at least one assessed domain, but overall quality ratings were relatively low in many articles. The Q-SSP, a tool which was deemed appropriate for our purposes, was designed specifically for survey studies but did not provide clear guidance on rating the reporting quality of studies using secondary institutional data. This made scoring more difficult for certain items and required extensive discussion related to data collection and analyses practices (i.e., handling of missing data) and ethics (i.e., consent and debriefing) as this information is generally viewed as an indicator of study quality (Appelbaum et al., 2018). Missing data can impact the interpretation and generalizability of study findings (Rubin, 2009), as we observed in studies examining the non-response bias on SRI results (Bacon et al., 2016; Goos & Salomons, 2017; Macfadyen et al., 2016; Treischl & Wolbring, 2017). However, although quality of reporting can serve as a proxy of the quality of the studies, these are not the same thing. Low reporting quality overall does not mean unreliable data, inappropriate analyses, or incorrect conclusions. In our review, we found evidence that analyses were justified, and reporting was appropriately discussed in terms of the study population. Regardless, to strengthen research reports on SRIs, authors must adhere to reporting standards (Appelbaum et al., 2018). Future work might also include the use of quality assessment measures for education studies that examine secondary data.

Conclusions

The research examined in this systematic review identified various sources of bias that have a meaning-

ful, cumulative effect on instructor and course evaluations. Results from these evaluations often have a negative impact that cannot be ignored. Our findings highlight the importance of mitigating biases prior to SRI completion and interpreting student ratings with caution. Despite the overwhelming evidence of bias in SRIs, institutions continue to implement SRIs. If SRIs continue to be used, faculty and administrators should consider implementing various evidence-informed practices to reduce the likelihood of bias and reducing the impact of biased results and understand the role of higher education policies for EDI in evaluating teaching. We suggest that attempts to control for biases in the data after collection, by calculating corrected means or matching courses and instructors across many dimensions, should be avoided as they may not be feasible and may introduce new sources of bias. However, if used in conjunction with other evaluation methods, SRIs may provide insight into areas for teaching development, and should not be used as a lone measure to make hiring, promotion, and tenure decisions.

References

*Indicates articles that were reviewed in this systematic review.

*Al-Maamari, F. (2015). Response rate and teaching effectiveness in institutional student evaluation of teaching: A multiple linear regression study. *Higher Education Studies, 5*(6), 9–20. https://www.ccsenet.org/journal/index.php/hes

*Alauddin, M., & Kifle, T. (2014). Does the student evaluation of teaching instrument really measure instructors' teaching effectiveness? An econometric analysis of students' perceptions in economics courses. *Economic Analysis and Policy, 44*(2), 156–168. https://doi.org/10.1016/j.eap.2014.05.009

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and Communications Board task force report. *American Psychologist, 73*(1), 3–25. https://doi.org/10.1037/amp0000191

*Arnold, I. J. M., & Versluis, I. (2019). The influence of cultural values and nationality on student evaluation of teaching. *International Journal of Educational Research, 98*, 13–24. https://doi.org/10.1016/j.ijer.2019.08.009

Arreola, R. A. (2007). *Developing a comprehensive faculty evaluation system: A guide to designing, building, and operating large-scale faculty evaluation systems.* Anker Publishing Company.

*Arrona-Palacios, A., Okoye, K., Camacho-Zuniga, C., Hammout, N., Luttmann-Nakamura, E., Hosseini, S., & Escamilla, J. (2020). Does professors' gender impact how students evaluate their teaching and the recommendations for the best professor? *HELIYON, 6*(10). https://doi.org/10.1016/j.heliyon.2020.e05313

*Bacon, D. R., Johnson, C. J., & Stewart, K. A. (2016). Nonresponse bias in student evaluations of teaching. *Marketing Education Review, 26*(2), 93–104. https://doi.org/10.1080/10528008.2016.1166442

*Bahous, S. A., Salameh, P., Salloum, A., Salameh, W., Park, Y. S., & Tekian, A. (2018). Voluntary vs. compulsory student evaluation of clerkships: effect on validity and potential bias. *BMC Medical Education, 18*. https://doi.org/10.1186/s12909-017-1116-8

Becker, W. E., & Watts, M. (1999). How departments of economics should evaluate teaching. *American Economic Review, 89*(2), 344–349. http://dx.doi.org/10.1257/aer.89.2.344

Benton, S. L., & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. In M. B. Paulsen (Ed.), *Higher education: Handbook of theory & research* (Vol. 29, pp. 279–326). Springer. https://doi.org/10.1007/978-94-017-8005-6_7

Benton, S. L., & Ryalls, K. R. (2016). Challenging misconceptions about student ratings of instruction. *The IDEA Center, 58*, 1-22. https://files.eric.ed.gov/fulltext/ED573670.pdf

*Bianchini, S., Lissoni, F., & Pezzoni, M. (2013). Instructor characteristics and students' evaluation of teaching effectiveness: Evidence from an Italian engineering school. *European Journal of Engineering Education, 38*(1), 38–57. https://doi.org/10.1080/03043797.2012.742868

*Blecich, A. A., & Zaninović, V. (2019). Insight into students' perception of teaching: Case of economic higher education instititution. *Journal of Contemporary Management Issues, 24*(1), 137–152. https://doi.org/10.30924/mjcmi.24.19

*Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics, 145*, 27–41. https://doi.org/10.1016/j.jpubeco.2016.11.006

*Borkan, B. (2017). Exploring variability sources in student evaluation of teaching via many-facet Rasch model. *Journal of Measurement and Evaluation in Education and Psychology, 8*(1), 15–33. https://doi.org/10.21031/epod.298462

Braskamp, L. A. ., & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. Jossey-Bass Publishers.

Carrell, S. E., Page, M. E., & West, J. E. (2010). Sex and science: How professor gender perpetuates the gender gap. *Quarterly Journal of Economics, 125*(3), 1101–1144. https://doi.org/10.1162/qjec.2010.125.3.1101

Centra, J. A. (1976). The influence of different directions on student ratings of instruction. *Journal of Educational Measurement, 13*(4), 277–282. https://www.jstor.org/stable/1434104

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. Jossey-Bass.

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

Centra, J. A. (2009). *Differences in responses to the student instructional report: Is it bias?* Princeton: Educational Testing Service.

Centre for Teaching Support & Innovation. (2018). *University of Toronto's Cascaded Course Evaluation Framework: Validation Study of the Institutional Composite Mean (ICM).*

*Chávez, K. (2020). Exploring bias in student evaluations: Gender, race, and ethnicity. *PS: Political Science & Politics, 53*(2), 270–274. https://doi.org/10.1017/S1049096519001744

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn?: A meta-analysis and review of the literature. *Journal of Marketing Education, 31*(1), 16–30. https://doi.org/10.1177/0273475308324086

*Dodeen, H. (2013). Validity, reliability, and potential bias of short rorms of students' evaluation of teaching: The case of UAE university. *Educational Assessment, 18*(4), 235–250. https://doi.org/10.1080/10627197.2013.846670

*Esarey, J., & Valdes, N. (2020). Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment and Evaluation in Higher Education, 45*(8), 1106–1120. https://doi.org/10.1080/02602938.2020.1724875

*Estelami, H. (2015). The effects of survey timing on student evaluation of teaching measures obtained using online surveys. *Journal of Marketing Education, 37*(1), 54–64. https://doi.org/10.1177/0273475314552324

*Ewing, A. M. (2012). Estimating the impact of relative expected grade on student evaluations of teachers. *Economics of Education Review, 31*(1), 141–154. https://doi.org/10.1016/j.econedurev.2011.10.002

Fairlie, R. W., Hoffmann, F., & Oreopoulos, P. (2014). A community college instructor like me: Race and ethnicity interactions in the classroom. *American Economic Review, 104*(8), 2567–2591. https://doi.org/10.1257/aer.104.8.2567

*Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019). Gender and cultural bias in student evaluations: Why representation matters. *Public Library of Science One, 14*(2). https://doi.org/10.1371/journal.pone.0209749

*Fassiotto, M., Li, L., Maldonado, Y., & Kothary, N. (2018). Female surgeons as counter stereotype: The impact of gender perceptions on trainee evaluations of physician faculty. *Journal of Surgical Education, 75*(5), 1140–1148. https://doi.org/10.1016/j.jsurg.2018.01.011

*Feistauer, D., & Richter, T. (2018a). The role of clarity about study programme contents and interest in student evaluations of teaching. *Psychology Learning and Teaching, 17*(3), 272–292. https://doi.org/10.1177/1475725718779727

*Feistauer, D., & Richter, T. (2018b). Validity of students' evaluations of teaching: Biasing effects of likability and prior subject interest. *Studies in Educational Evaluation*, *59*, 168–178. https://doi.org/10.1016/j.stueduc.2018.07.009

Ferguson-Patrick, K. (2011). Professional development of early career teachers: A pedagogical focus on cooperative learning. *Issues in Educational Research*, *21*(2), 109–129. http://iier.org.au/iier21/ferguson-patrick.html

*Fischer, E., & Hänze, M. (2019). Bias hypotheses under scrutiny: Investigating the validity of student assessment of university teaching by means of external observer ratings. *Assessment & Evaluation in Higher Education*, *44*(5), 772–786. https://doi.org/10.1080/02602938.2018.1535647

*Flegl, M., & Andrade Rosas, L. A. (2019). Do professor's age and gender matter or do students give higher value to professors' experience? *Quality Assurance in Education: An International Perspective*, *27*(4), 511–532. https://doi.org/10.1108/QAE-12-2018-0127

*Fogarty, T. J., Jonas, G. A., & Parker, L. M. (2013). The medium is the message: Comparing paper-based and web-based course evaluation modalities. *Journal of Accounting Education*, *31*(2), 177–193. https://doi.org/10.1016/j.jaccedu.2013.03.002

Galbraith, C., Merrill, G., & Kline, D. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and Bayesian analyses. *Research in Higher Education*, *53*(3), 353–374. https://doi.org/10.1007/s11162-011-9229-0

*Gith, E. (2020). The impact of the Israeli-Palestinian conflict on thinking biases in teaching evaluations. *Peace and Conflict: Journal of Peace Psychology*, *26*(1), 92–95. https://doi.org/10.1037/pac0000386

*Goos, M., & Salomons, A. (2017). Measuring teaching quality in higher education: Assessing selection bias in course evaluations. *Research in Higher Education*, *58*(4), 341–364. https://doi.org/10.1007/s11162-016-9429-8

Gravestock, P., & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models and trends*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.627.559&rep=rep1&type=pdf

*Griffin, T. J., Hilton III, J., Plummer, K., & Barret, D. (2014). Correlation between grade point averages and student evaluation of teaching scores: Taking a closer look. *Assessment & Evaluation in Higher Education*, *39*(3), 339–348. https://doi.org/10.1080/02602938.2013.831809

*Gupta, A., Garg, D., & Kumar, P. (2018). Analysis of students' ratings of teaching quality to understand the role of gender and socio-economic diversity in higher education. *IEEE Transactions on Education*, *61*(4), 319–327. https://doi.org/10.1109/TE.2018.2814599

Heffernan, T. (2021). Sexism, racism, prejudice, and bias: A literature review and synthesis of research surrounding student evaluations of courses and teaching. *Assessment & Evaluation in Higher Education*, *0*(0), 1–11. https://doi.org/10.1080/02602938.2021.1888075

Hoffman, F., & Oreopoulos, P. (2009). A professor like me: The influence of instructor gender on college achievement. *Journal of Human Resources*, *44*(2), 479–494. https://doi.org/10.3368/jhr.44.2.479

Hofstede, G. (1986). Culutral differences in teaching and learning. *International Journal of Intercultural Relations*, *10*, 301–320. https://doi.org/10.1016/0147-1767(86)90015-5

*Jobu Babin, J., Hussey, A., Nikolsko-Rzhevskyy, A., & Taylor, D. A. (2020). Beauty premiums among academics. *Economics of Education Review*, *78*, 102019. https://doi.org/10.1016/j.econedurev.2020.102019

*Laupper, E., Balzer, L., & Berger, J.-L. (2020). Online vs. offline course evaluation revisited: testing the invariance of a course evaluation questionnaire using a multigroup confirmatory factor analysis framework. *Educational Assessment, Evaluation and Accountability*, *32*(4), 481–498. https://doi.org/10.1007/s11092-020-09336-6

Lindqvist, A., Sendén, M. G., & Renström, E. A. (2020). What is gender, anyway: A review of the options for operationalising gender. *Psychology and Sexuality*, 1–13. https://doi.org/10.1080/19419899.2020.1729844

Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation*, *54*, 94–106. https://doi.org/10.1016/j.stueduc.2016.12.004

*Liu, O. L. (2012). Student evaluation of instruction: In the new paradigm of distance education. *Research in Higher Education*, *53*(4), 471–486. https://doi.org/10.1007/s11162-011-9236-1

Llamas, J. D., Nguyen, K., & Tran, A. G. T. T. (2021). The case for greater faculty diversity: examining the educational impacts of student-faculty racial/ethnic match. *Race Ethnicity and Education*, *24*(3), 375–391. https://doi.org/10.1080/13613324.2019.1679759

Louie, D. W., Poitras-Pratt, Y., Hanson, A. J., Ottmann, J. (2017). Applying Indigenizing Principles of Decolonizing Methodologies in University Classrooms. *Canadian Journal of Higher Education/Revue canadienne d'enseignement supérieur*, *47*(3), 16–33. https://doi.org/10.7202/1043236ar

*Macfadyen, L. P., Dawson, S., Prest, S., & Gaševic, D. (2016). Whose feedback? A multilevel analysis of student completion of end-of-term teaching evaluations. *Assessment & Evaluation in Higher Education*, *41*(6), 821–839. https://doi.org/10.1080/02602938.2015.1044421

*Magel, R. C., Doetkott, C., & Cao, L. (2017). A study of the relationship between gender, salary, and student ratings of instruction at a research university. *NASPA Journal About Women in Higher Education*, *10*(1), 96–117. https://doi.org/10.1080/19407882.2017.1285792

*Maricic, M., Dokovic, A., & Jeremic, V. (2019). The validity of student evaluation of teaching: Is there a gender bias? *Croatian Journal of Education*, *21*(3), 743–775. https://doi.org/10.15516/cje.v21i3.3177

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Springer.

*Martin, L. L. (2016). Gender, teaching evaluations, and professional success in political science. *PS: Political Science & Politics*, *49*(2), 313–319. https://doi.org/10.1017/S1049096516000275

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

Medina, M. S., Smith, T., Kolluru, S., Sheaffer, E. A., & ViVall, M. (2019). A review of strategies for designing, administering, and using student ratings of instruction. *American Journal of Pharmaceutical Education*, *83*(5), 753–764. https://doi.org/10.5688/ajpe7177

*Mitchell, K. M. W., & Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science & Politics*, *51*(3), 648–652. https://doi.org/10.1017/S104909651800001X

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Physical Therapy*, *89*(9), 873–880. https://doi.org/10.1136/bmj.b2535

*Nargundkar, S., & Shrikhande, M. (2014). Norming of student evaluations of instruction: Impact of noninstructional factors. *Decision Sciences Journal of Innovative Education*, *12*(1), 55–72. https://scholarworks.gsu.edu/managerialsci_facpub/3

Nicolaou, M., & Atkinson, M. (2019). Do student and survey characteristics affect the quality of UK undergraduate medical education course evaluation? A systematic review of the literature. *Studies in Educational Evaluation*, *62*, 92–103. https://doi.org/10.1016/j.stueduc.2019.04.011

*Okoye, K., Arrona-Palacios, A., Camacho-Zuniga, C., Hammout, N., Nakamura, E. L., Escamilla, J., & Hosseini, S. (2020). Impact of students evaluation of teaching: A text analysis of the teachers qualities by gender. *International Journal of Educational Technology in Higher Education*, *17*(1). https://doi.org/10.1186/s41239-020-00224-z

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan-a web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 1–10. https://doi.org/10.1186/s13643-016-0384-4

*Palali, A., van Elk, R., Bolhaar, J., & Rud, I. (2018). Are good researchers also good teachers? The relationship between research quality and teaching quality. *Economics of Education Review*, *64*, 40–49. https://doi.org/10.1016/j.econedurev.2018.03.011

*Park, E., & Dooris, J. (2020). Predicting student evaluations of teaching using decision tree analysis. *Assessment & Evaluation in Higher Education*, *45*(5), 776–793. https://doi.org/10.1080/02602938.2019.1697798

*Park, H.-S., & Cheong, Y. F. (2018). Correlates of monotonic response patterns in online ratings of a university course. *Higher Education*, *76*(1), 101–113. https://doi.org/10.1007/s10734-017-0199-9

*Peterson, D. A. M., Biederman, L. A., Andersen, D., Ditonto, T. M., & Roe, K. (2019). Mitigating gender bias in student evaluations of teaching. *Public Library of Science One*, *14*(5).

*Protogerou, C., & Hagger, M. S. (2020). A checklist to assess the quality of survey studies in psychology. *Methods in Psychology*, *3*(July), 100031. https://doi.org/10.1016/j.metip.2020.100031

*Punyanunt-Carter, N., & Carter, S. L. (2015). Students' gender bias in teaching evaluations. *Higher Learning Research Communications*, *5*(3), 28. https://doi.org/10.18870/hlrc.v5i3.234

*Radchenko, N. (2020). Student evaluations of teaching: Unidimensionality, subjectivity, and biases. *Education Economics*, *28*(6), 549–566. https://doi.org/10.1080/09645292.2020.1814997

Ray, B., Babb, J., & Wooten, C. A. (2018). Rethinking SETs: Retuning student evaluations of teaching for student agency. *Composition Studies*, *46*(1), 34–56. https://compstudiesjournal.com/46-1/

*Reisenwitz, T. H. (2016). Student evaluation of teaching: An investigation of nonresponse bias in an online context. *Journal of Marketing Education*, *38*(1), 7–17. https://doi.org/10.1177/0273475315596778

*Risquez, A., Vaughan, E., & Murphy, M. (2015). Online student evaluations of teaching: what are we sacrificing for the affordances of technology? *Assessment & Evaluation in Higher Education*, *40*(1), 120–134. https://doi.org/10.1080/02602938.2014.890695

*Rodríguez, A. M., Capelleras, J.-L., & Garcia, V. M. G. (2014). Teaching performance: Determinants of the student assessment. *Academia*, *27*(3), 402–418. https://doi.org/10.1108/ARLA-11-2013-0177

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

*Royal, K. D., & Stockdale, M. R. (2015). Are teacher course evaluations biased against faculty that teach quantitative methods courses? *International Journal of Higher Education*, *4*(1), 217–224. www.sciedu.ca/ijhe

Rubin, D. B. (2009). *Multiple imputation for nonresponse in surveys*. WILEY.

Schiekirka, S., & Raupach, T. (2015). A systematic review of factors influencing student ratings in undergraduate medical education course evaluations. *BMC Medical Education*, *15*(1), 1–9. https://doi.org/10.1186/s12909-015-0311-8

*Schönrock-Adema, J., Lubarsky, S., Chalk, C., Steinert, Y., & Cohen-Schotanus, J. (2013). "What would my classmates say?" An international study of the prediction-based method of course evaluation. *Medical Education*, *47*(5), 453–462. https://doi.org/10.1111/medu.12126

*Schueths, A. M., Gladney, T., Crawford, D. M., Bass, K. L., & Moore, H. A. (2013). Passionate pedagogy and emotional labor: Students' responses to learning diversity from diverse instructors. *International Journal of Qualitative Studies in Education*, *26*(10), 1259–1276. https://doi.org/10.1080/09518398.2012.731532

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & Group, P.-P. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ (Clinical Research Ed.), 349*(3), g7647–g7647. https://doi.org/10.1136/bmj.g7647

*Socha, A. (2013). A hierarchical approach to students' assessments of instruction. *Assessment & Evaluation in Higher Education*, *38*(1), 94–113. https://doi.org/10.1080/02602938.2011.604713

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

*Spooren, P., & Christiaens, W. (2017). I liked your course because I believe in (the power of) student evaluations of teaching (SET). Students' perceptions of a teaching evaluation process and their relationships with SET scores. *Studies in Educational Evaluation*, *54*, 43–49. https://doi.org/10.1016/j.stueduc.2016.12.003

Stroebe, W. (2016). Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations. *Perspectives on Psychological Science*, *11*(6), 800–816. https://doi.org/10.1177/1745691616650284

*Sulis, I., Porcu, M., & Capursi, V. (2019). On the use of student evaluation of teaching: A longitudinal analysis combining measurement issues and implications of the exercise. *Social Indicators Research*, *142*(3), 1305–1331. https://doi.org/10.1007/s11205-018-1946-8

*Tarun, P., & Krueger, D. (2016). A perspective on student evaluations, teaching techniques, and critical thinking. *Journal of Learning in Higher Education*, *12*(2), 1–13. https://www.jwpress.com/

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

*Tomes, T., Coetzee, S., & Schmulian, A. (2019). Prediction-based student evaluations of teaching as an alternative to traditional opinion-based evaluations. *Assessment & Evaluation in Higher Education*, *44*(8), 1222–1236. https://doi.org/10.1080/02602938.2019.1594157

*Treischl, E., & Wolbring, T. (2017). The causal effect of survey mode on students' evaluations of teaching: Empirical evidence from three field experiments. *Research in Higher Education*, *58*(8), 904–921. https://doi.org/10.1007/s11162-017-9452-4

*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

*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

*Wang, L., & Gonzalez, J. A. (2020). Racial/ethnic and national origin bias in SET. *International Journal of Organizational Analysis*, *28*(4), 843–855. https://doi.org/10.1108/IJOA-06-2019-1793

*Weidman-Evans, E., Hayes, S., & Bigler, T. (2020). Relationship between course evaluations and course grades in six allied health programs. *Health Professions Education*, *6*(4), 612–616. https://doi.org/10.1016/j.hpe.2020.07.006

*Winer, L., DiGenova, L., & Costopoulos, A. (2016). Addressing common concerns about online student ratings of instruction: A research-informed approach. *Canadian Journal of Higher Education*, *46*(4), 115–131. https://journals.sfu.ca/cjhe/index.php/cjhe/article/view/186112/pdf

Wolbring, T. (2012). Class attendance and students' evaluations of teaching: Do no-shows bias course ratings and rankings? *Evaluation Review*, *36*(1), 72–96. https://doi.org/10.1177/0193841X12441355

*Wolbring, T., & Treischl, E. (2016). Selection bias in students' evaluation of teaching: Causes of student absenteeism and its consequences for course ratings and rankings. *Research in Higher Education, 57*(1), 51–71. https://doi.org/10.1007/s11162-015-9378-7

Wright, S. L., & Jenkins-Guarnieri, M. A. (2012). Student evaluations of teaching: Combining the meta-analyses and demonstrating further evidence for effective use. *Assessment and Evaluation in Higher Education, 37*(6), 683–699. https://doi.org/10.1080/02602938.2011.563279

*Yueh, H.-P., Chen, T.-L., Chiu, L.-A., Lee, S.-L., & Wang, A.-B. (2012). Student evaluation of teaching effectiveness of a nationwide innovative education program on image display technology. *IEEE Transactions on Education, 55*(3), 365–369. https://doi.org/10.1109/TE.2011.2178121