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)

<table>
<thead>
<tr>
<th>First Author Last Name (Year); Country of study</th>
<th>Number of SRIs</th>
<th>Students</th>
<th>Evaluated Instructors</th>
<th>Type of bias: Summary of findings</th>
<th>Reporting Quality Assessment Domain Scores (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanuddin (2014); Australia</td>
<td>10,223</td>
<td>NR</td>
<td>undergrad, graduate</td>
<td>Personality, culture: Higher SRIs associated with instructor characteristics (e.g., organization, expertise, enthusiasm, helpfulness, respectfulness) and English (vs non-English) speaking background.</td>
<td>75.0 33.3 33.3 0</td>
</tr>
<tr>
<td>Al-Maamari (2015); Oman</td>
<td>NR</td>
<td>2,095</td>
<td>English language program</td>
<td>Gender, class size, course type, nonresponse: Female (vs male) instructors and elective (vs required) courses received lower SRIs.</td>
<td>75.0 66.7 77.8 0</td>
</tr>
<tr>
<td>Arnold (2019); Netherlands</td>
<td>NR</td>
<td>765</td>
<td>M = 21.5</td>
<td>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.</td>
<td>75.0 33.3 55.6 0</td>
</tr>
<tr>
<td>Arroño-Palacios (2020); Mexico</td>
<td>NR</td>
<td>103,833</td>
<td>NR</td>
<td>Gender: Students more likely to report male (vs female) professors as their best professors.</td>
<td>75.0 66.7 44.4 100</td>
</tr>
<tr>
<td>Bacon (2016); USA</td>
<td>6,754</td>
<td>NR</td>
<td>NR</td>
<td>Nonresponse: Low response rates advantage instructors with high SRIs and disadvantage instructors with low SRIs.</td>
<td>100 33.3 66.7 0</td>
</tr>
<tr>
<td>Bahous (2018); Lebanon</td>
<td>NR</td>
<td>363</td>
<td>41% F, 59% M; 3rd year medical students</td>
<td>SRI administration procedures: Compulsory SRIs did not improve reliability or influence results but were linked to increased inattentive responding rates.</td>
<td>50.0 66.7 66.7 100</td>
</tr>
<tr>
<td>Bianchini (2013); France</td>
<td>NR</td>
<td>1,756</td>
<td>undergrad, graduate</td>
<td>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.</td>
<td>75.0 66.7 66.7 0</td>
</tr>
<tr>
<td>Blečić (2019); Croatia</td>
<td>NR</td>
<td>333</td>
<td>NR</td>
<td>Interest in course content, grades, class size: SRIs positively influenced by students’ grade and (to a lesser extent) class size.</td>
<td>75.0 33.3 66.7 0</td>
</tr>
<tr>
<td>Boring (2017); France</td>
<td>NR</td>
<td>4,362</td>
<td>M = 18</td>
<td>Gender: Male students gave higher SRIs to male (vs female) professors, even when student grades are considered.</td>
<td>75.0 66.7 55.6 100</td>
</tr>
<tr>
<td>Borkan (2017); Turkey</td>
<td>NR</td>
<td>1,235</td>
<td>NR</td>
<td>Course type expected grade: Higher SRIs from students in elective (vs required) courses and with higher (vs lower) expected grades.</td>
<td>25.0 33.3 55.6 100</td>
</tr>
<tr>
<td>Chávez (2020); US</td>
<td>NR</td>
<td>42</td>
<td>NR</td>
<td>Gender, ethnicity: Female instructors and instructors of colour receive lower scores than their male and white counterparts.</td>
<td>75.0 66.7 55.6 0</td>
</tr>
<tr>
<td>Study, Year, Country</td>
<td>Sample Size</td>
<td>Gender, Age, Experience, Course Type</td>
<td>SRI Item Phrasing</td>
<td>Study Design, Instructor Bias, SRI Association</td>
<td>Bias Against Female Instructors, International Students</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------</td>
<td>------------------------------------</td>
<td>------------------</td>
<td>-----------------------------------------------</td>
<td>-----------------------------------------------------</td>
</tr>
<tr>
<td>Dodeen (2013), UAE</td>
<td>NR 3,661</td>
<td>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.</td>
<td>NR</td>
<td>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.</td>
<td>Nonresponse, response timing: Late responders gave lower SRIs.</td>
</tr>
<tr>
<td>Author</td>
<td>Year</td>
<td>Country</td>
<td>Sample Size</td>
<td>Gender Distribution</td>
<td>Methods/Results</td>
</tr>
<tr>
<td>-----------------</td>
<td>------</td>
<td>-------------</td>
<td>-------------</td>
<td>---------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Fogarty</td>
<td>2013</td>
<td>US</td>
<td>NR</td>
<td>NR</td>
<td>SRI administration procedures: Lower SRIs produced with web- (vs paper-) based surveys.</td>
</tr>
<tr>
<td>Gith</td>
<td>2020</td>
<td>Israel</td>
<td>36,712</td>
<td>27,422 Jewish, 9,290 Arab</td>
<td>Ethnicity/culture: Students rated instructors who are members of their own cultural group higher.</td>
</tr>
<tr>
<td>Goos</td>
<td>2017</td>
<td>European</td>
<td>NR</td>
<td>NR</td>
<td>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.</td>
</tr>
<tr>
<td>Gith</td>
<td>2020</td>
<td>Israel</td>
<td>NR</td>
<td>NR</td>
<td>GPA: Lower GPA associated with higher SRIs.</td>
</tr>
<tr>
<td>Griffin</td>
<td>2014</td>
<td>US</td>
<td>2,073</td>
<td>NR</td>
<td>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.</td>
</tr>
<tr>
<td>Gupta</td>
<td>2018</td>
<td>India</td>
<td>112,919</td>
<td>NR</td>
<td>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.</td>
</tr>
<tr>
<td>Jobu Babin</td>
<td>2020</td>
<td>US</td>
<td>2,968</td>
<td>NR</td>
<td>Level of study, course type, rank, gender, class size: First year (vs higher level) students provide least 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.</td>
</tr>
<tr>
<td>Laupper</td>
<td>2020</td>
<td>Switzerland</td>
<td>463</td>
<td>NR</td>
<td>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.</td>
</tr>
<tr>
<td>Liu</td>
<td>2012</td>
<td>US</td>
<td>11,351</td>
<td>NR</td>
<td>Gender: Male salaries increase as SRIs increase. Female salaries decrease as SRIs increase.</td>
</tr>
<tr>
<td>Macfayden</td>
<td>2016</td>
<td>Canada</td>
<td>21,534</td>
<td>NR</td>
<td>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.</td>
</tr>
<tr>
<td>Magel</td>
<td>2017</td>
<td>US</td>
<td>NR</td>
<td>NR</td>
<td>Gender: Male salaries increase as SRIs increase. Female salaries decrease as SRIs increase.</td>
</tr>
<tr>
<td>Name</td>
<td>Year</td>
<td>Sample Size</td>
<td>Gender</td>
<td>Course Level</td>
<td>Instructor Rank</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-------</td>
<td>-------------</td>
<td>--------</td>
<td>--------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Maricic (2019); Croatia</td>
<td></td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>6</td>
</tr>
<tr>
<td>Martin (2016); US</td>
<td>309</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Mitchell (2018); US</td>
<td>1,424</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>2</td>
</tr>
<tr>
<td>Nargundkar (2014); US</td>
<td>NR</td>
<td>105,974</td>
<td>NR</td>
<td>undergrad, graduate</td>
<td>NR</td>
</tr>
<tr>
<td>Okoye (2020); Mexico</td>
<td>82,144</td>
<td>NR</td>
<td>NR</td>
<td>undergrad</td>
<td>NR</td>
</tr>
<tr>
<td>Palali (2018); Netherlands</td>
<td>28,243</td>
<td>9,000</td>
<td>M = 22</td>
<td>44% F, 56% M; 48% undergrad, 52% graduate</td>
<td>83</td>
</tr>
<tr>
<td>Park (2020); US</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>17% first-year, 16% sophomore, 24% junior, 23% senior, 16% graduate</td>
<td>2,870</td>
</tr>
<tr>
<td>Park, H-S (2018); Korea</td>
<td></td>
<td>1,206</td>
<td>NR</td>
<td>42% F, 58% M; 21% Year 1, 79% upper level</td>
<td>NR</td>
</tr>
<tr>
<td>Peterson (2019); US</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>4</td>
</tr>
<tr>
<td>Punyanunt-Carter (2015); US</td>
<td></td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Study</td>
<td>Country</td>
<td>Sample Size</td>
<td>Gender Distribution</td>
<td>Course Type</td>
<td>Class Size</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------------</td>
<td>-------------</td>
<td>---------------------</td>
<td>-------------</td>
<td>------------</td>
</tr>
<tr>
<td>Radchenko (2020); US</td>
<td>NR 365,187</td>
<td>61% F, 39% M; undergrad, graduate</td>
<td>2,093</td>
<td>40% F, 60% M; 60% full time professors</td>
<td>Gender, course type, class size: SRIIs 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 SRIIs.</td>
</tr>
<tr>
<td>Reisenwitz (2016); US</td>
<td>NR 313</td>
<td>M = 20.6 42.5% F, 57.5% M; 60% White; undergrad</td>
<td>NR</td>
<td>NR</td>
<td>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.</td>
</tr>
<tr>
<td>Risquez (2015); Ireland</td>
<td>NR 63,173</td>
<td>NR NR NR 673</td>
<td>NR</td>
<td>NR</td>
<td>SRI administration procedures, class size, preparation: Delivery mode (paper vs online) had no effect on SRIs after controlling for class size, faculty.</td>
</tr>
<tr>
<td>Rodríguez (2014); Spain</td>
<td>NR 1,359</td>
<td>M = 20.3 57.1% men, 42.9% women</td>
<td>125</td>
<td>NR</td>
<td>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.</td>
</tr>
<tr>
<td>Royal (2015); US</td>
<td>NR 2,564</td>
<td>NR NR NR NR</td>
<td>NR</td>
<td>NR</td>
<td>Course type: Students were more critical of instructors of methods (vs non-methods) courses. 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.</td>
</tr>
<tr>
<td>Schönrock-Adema (2013); Canada, Netherlands</td>
<td>NR 966</td>
<td>NR NR NR NR</td>
<td>NR</td>
<td>NR</td>
<td>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.</td>
</tr>
<tr>
<td>Schueths (2013); US</td>
<td>NR NR NR NR 29</td>
<td>M = 42.6, M = 46.6</td>
<td>20.6% ethnic minority, 79.3% white; 6 teaching assistants, 10 lecturers, 10 professors, 4 other</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Socha (2013); US</td>
<td>NR 4,063</td>
<td>undergrad, graduate 89</td>
<td>NR</td>
<td>NR</td>
<td>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.</td>
</tr>
<tr>
<td>Author</td>
<td>Year</td>
<td>Country</td>
<td>Sample Size</td>
<td>Gender</td>
<td>Ethnicity</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>--------</td>
<td>-----------</td>
</tr>
<tr>
<td>Spooren</td>
<td>2017</td>
<td>Belgium</td>
<td>927</td>
<td>66% F, 34% M; undergrad</td>
<td>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 life science students value SRIs more, leading to higher SRIs.</td>
</tr>
<tr>
<td>Sulis</td>
<td>2019</td>
<td>Italy</td>
<td>6,425</td>
<td>65.5% F, 34.5% M</td>
<td>Interest in course content, instructor rank: SRIs influenced positively by students' prior interest and knowledge of course content. SRIs not associated with instructor rank.</td>
</tr>
<tr>
<td>Tarun</td>
<td>2016</td>
<td>US</td>
<td>209</td>
<td>63% F, 37% M, African, mixed-race, Indian, Asian, White; undergrad</td>
<td>Test type, grades, interest in course content, course difficulty/workload: SRIs influenced by assessment type, grades, interest in course, course difficulty, and student workload.</td>
</tr>
<tr>
<td>Tomas</td>
<td>2019</td>
<td>South Africa</td>
<td>257</td>
<td>63% F, 37% M</td>
<td>Gender, inattentive responses, nonresponse, ethnicity/culture, academic performance: Male (vs female) students and those with higher academic performance gave higher SRIs.</td>
</tr>
<tr>
<td>Treischl</td>
<td>2017</td>
<td>Germany</td>
<td>2,037</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Valencia</td>
<td>2020</td>
<td>Canada</td>
<td>3,000</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Wagner</td>
<td>2016</td>
<td>Netherlands</td>
<td>NR</td>
<td>93</td>
<td>NR</td>
</tr>
<tr>
<td>Wang</td>
<td>2020</td>
<td>US</td>
<td>264</td>
<td>71% M, 29% F</td>
<td>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.</td>
</tr>
<tr>
<td>Weidman-Evans</td>
<td>2020</td>
<td>US</td>
<td>18,669</td>
<td>43% F, 57% M; 33% non-Caucasian</td>
<td>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.</td>
</tr>
<tr>
<td>Winer</td>
<td>2016</td>
<td>Canada</td>
<td>18,000</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Wolfbring</td>
<td>2012</td>
<td>Germany</td>
<td>18,000</td>
<td>NR</td>
<td>NR</td>
</tr>
</tbody>
</table>
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.

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.

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).

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 & Chris-tiaens, 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 & Häenze, 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 & Salomon, 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.


Stoesz et al.  