Ethnic and Gender Biases in Clinical Performance Assessment (CPA) in Healthcare Education: A Systematic Review

Iris C.I. Chao (chengin@ualberta.ca)
University of Alberta https://orcid.org/0000-0002-9464-5840
Efrem Violato
University of Alberta
Brendan Concannon
University of Alberta
Charlotte McCartan
University of Alberta
Katarzyna Nicpon
University of Alberta
Sharla King
University of Alberta
Mary Roduta Roberts
University of Alberta

Research article

Keywords: clinical performance, assessment, OSCE, bias, ethnicity, gender, standardized patient

Posted Date: November 22nd, 2019

DOI: https://doi.org/10.21203/rs.2.17622/v1

Abstract

Background: Several forms of bias, including ethnic and gender bias, are thought to impact evaluations on Clinical Performance Assessments (CPAs). Unfairness may influence student learning attitudes if a loss of trust causes a lack of engagement in learning. Understanding the biases occurring in CPAs can lead to well-designed examiner training to ensure equality and fairness. The purpose of this systematic review is to determine the current evidence in the literature for ethnic and/or gender bias by examiners evaluating pre-licensure healthcare students in CPAs using standardized patients (SPs).

Methods: Literature was systematically searched in CINAHL, PubMed and Medline from inception to February 2019, and no date range was set. Studies related to the investigation of ethnic and/or gender biases occurring in CPAs using SPs for examining health professions students were selected. A systematic review was conducted to assess the methodological quality and strength of evidence of relevant research and to identify if any potential ethnic and/or gender bias occurred in CPAs. The Guidelines for Critical Review were used to appraise the selected studies.

Results: Nine studies published from 2003 to 2017 were retrieved for review. Three studies met all the Guidelines for Critical Review quality criteria, indicating stronger evidence of their outcomes, two of the studies reported ethnic and/or gender bias existing in the CPAs. Overall, four studies found ethnic and/or gender bias in CPAs, but all study results had small effect sizes.

Conclusions: No systematic and consistent bias was found across the studies; nonetheless, the possibility of ethnic or gender bias by some examiners cannot be ignored. To minimize potential examiner bias, the investigation of Frame of Reference training, multiple examiners per station, and combination assessments in CPAs is recommended.

Background

Health profession education requires students to develop competence across different areas of practice [1]. To evaluate clinical skills during healthcare education, Clinical Performance Assessments (CPAs) are widely used [2]. Although the CPA is more effective at determining the holistic capacities of students than traditional multiple-choice examinations, examiner variability related to ethnicity and/or gender threatens the fairness of ratings [3,4].

Clinical Performance Assessment (CPA)

Evaluating the clinical performance of students in the healthcare field is challenging as students must be competent in various complex clinical skills [5,6]. To address this issue, medical educators have improved their examination methods by replacing traditional written examinations with clinical performance assessment (CPA) [6,7]. In the literature, these methods of evaluation have been referred to as a CPA, clinical performance examination (CPX), clinical performance evaluation (CPE), and performance-based assessment (PBA). For this systematic review, the term CPA will be used.

In the 1970s, the CPA was developed to help medical educators move beyond knowledge evaluation and evaluate the clinical competence of medical students [8–10]. CPAs can include the use of simulation and written or videotaped scenarios presented to students. Students are assessed based on the presentation of their professional perspectives and the design of treatment plans [11,12]. CPAs with simulation involves standardized patients (SPs). SPs are persons taught to portray patients with carefully developed medical histories. In the past decades, the use of SPs in CPAs has been increasing [13–16], as SPs provide various advantages beyond written or videotaped scenarios or using real patients.

Compared with evaluating clinical performance through written or videotaped scenarios, the use of SPs can help examiners better evaluate students' ability to perform real clinical skills and integrate knowledge learned in simulated situations to real-life scenarios [5]. SPs provide advantages over using real patients by creating a standardized and consistent presentation of clinical issues allowing student responses and scores to be comparable [17]. Additionally, the risk of discomforting real patients can be eliminated. Due to the advantages and the trend of using SPs in CPAs, this systematic review only includes studies that conducted CPAs with SPs. A common type of CPA involving SPs is the objective structured clinical examination (OSCE) [18,19].

Objective Structured Clinical Examination (OSCE)

In North America and Europe, the OSCE is the standard tool to help educators examine clinical competence of students in medicine, nursing, occupational therapy, physical therapy, dentistry, and pharmacy [20–24]. An OSCE consists of several tasks in which students are expected to perform a variety of clinical skills within a specified period when encountering SPs. The OSCE is useful in assessing students' clinical reasoning and communication skills, and students' abilities to take patients' medical histories and examine patients' functions [25,26]. Examiners typically use a rubric, rating scale, or checklist to rate students, where criteria scores and global ratings are assigned by examiners. Criteria scores refer to performance on a certain dimension, while global ratings describe the overall performance on an entire examination [25].

In the present review, most selected studies used the OSCE to evaluate clinical competence. OSCEs in different institutes or professions may have different designs. For instance, the number of dimensions and the time period of the assessment can vary. The naming of OSCEs is also variable. For example, two of the selected studies implemented the Practical Assessment of Clinical Skills in the Membership of the Royal Colleges of Physicians in the United Kingdom [MRCP (UK)—PACES] and its new version (nPACES) which are both OSCE-style assessments [3,27]. One of the selected studies used the Clinical Skills Assessment (CSA) for the Membership of the Royal College of General Practitioners (MRCGP), another OSCE-style assessment.

Many studies have reported that CPAs using SPs, such as the OSCE, are reliable and valid to evaluate the clinical performance of healthcare students, and allow examiners to provide more accurate feedback to students [28–34]. However, an issue for CPAs is the possible subjectivity of examiners that may introduce the risk of unfairness and bias, especially when examiners are affected by construct-irrelevant characteristics, such as ethnicity and/or gender [3,35–37].

Clinical performance of students based on ethnic minority status

Student populations in the health profession education have become more diverse in Western societies [38,39]. Numerous studies reported that students from ethnic minority groups perform less well than those from the ethnic majority in clinical performance [36,40–42] with communication skills tending to be the most significant factors affecting clinical performance [40,43,44]. Minority students often speak a second language when completing a CPA and so lower grades might be associated with their fluency and confidence in communication [40,45]. Moreover, potential ethnic-related stereotypes may cause examiner bias in CPAs [36,38].

Ethnic-related bias

Unconscious bias in CPA related to ethnicity can potentially influence student grades [39]. Woolf et al. [38] reported that possible ethnic-related stereotypes accounted for the underperformance of Asian medical students with examiners perceiving Asian students as possessing good professional knowledge but having poor communication skills and as not being active in class. A negative feedback loop may occur where students perceive clinical educators as having negative stereotypes of their ethnicity and in turn may become frustrated, in turn affecting their learning. Clinical educators with stereotypical views of students from ethnic minorities may have those stereotypes reinforced and feel less positive about teaching students of a certain ethnicity. Previous reports have theorized that African American visible minorities may over-identify their negative thoughts or emotions, resulting in a cycle of increasing anxiety symptomology [46–49].

Clinical performance of students based on gender

In addition to potential ethnic-related bias affecting student grades, the effect of gender on the clinical performance of students is also a major concern [50,51]. Female students are more likely to perform better on communication and interpersonal skills than males [52–55] as well as on the overall performance of both written tests and clinical skills [27,35,44,51,56]. Furthermore, studies reported that medical students who showed empathy received better clinical evaluations, with women receiving higher scores on empathy scales than men [57]. Possible gender bias occurs for males and females with equivalent interpersonal

abilities [52], where examiners expect females to have better interpersonal skills and as such are more lenient to female students while rating males of equivalent skill more harshly [58].

Gender-related bias

Research has found an interaction between examiners' and students' gender. One study found male medical students received better grades from female examiners than from male examiners in the OSCE, and female surgical students received higher marks from male examiners on the OSCE [35,51]. In contrast, Riese et al. [54] found that female examiners assigned higher scores to female students in a one-year clerkship CPA. Outside of the OSCE, some studies identified gender differences in grades for face-to-face oral examinations. Wiskin, Allan, and Skelkon [59] demonstrated that male-male examiner pairs awarded female students higher scores than male students compared to other examiner gender pairs. Niehaus, Jordaan, Koen, Mashile, and Mall [60] found that female-female examiner pairs were more likely to assign students lower marks compared to other examiner gender pairs in oral examinations. These studies suggest a likely gender bias in CPAs and oral examinations; however, the results were not consistent, some studies did not report any gender interaction effect in CPAs or oral examinations [27,50,56,61].

Purpose of the current systematic review

Although the practicality of utilizing SPs in CPAs is well-discussed in the literature, and the majority of studies report promising reliability and validity [62,63], other studies showed effects for examiner leniency and stringency, pre-assessment training content, and subjective bias [58,64,65]. These effects could represent a threat to inter-rater and intra-rater reliability and validity. In addition, unfairness may affect student learning attitudes if a loss of trust causes a lack of engagement in education. Understanding the biases occurring in CPAs can lead to well-designed examiner training to ensure equality and fairness. This study aims to systematically review the evidence for ethnic and/or gender bias by examiners evaluating pre-licensure healthcare students in CPAs using SPs.

Methods

This systematic review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [66].

Search strategy and eligibility criteria

Literature was searched in CINAHL, PubMed and Medline. We recruited studies published in English, and no date range was set. This purpose of this review is to investigate the evidence of ethnic and gender examiner biases occurring in CPAs using SPs when examining pre-licensure students. For finding relevant research, the search terms used were ("assessor bias" or "rater bias" or "examiner bias" or "bias") AND ("clinical performance" or "clinical competence" or "clinical skills" or "performance-based") AND ("racial" or "ethnic" or "race" or "ethnicity" or "gender" or "sex") AND ("education" or "healthcare" or "clinical"). Three authors (IC, EV, and CM) independently searched and screened the literature.

The inclusion criteria of this review were the literature (1) exploring examiner biases related to ethnicity and gender; (2) using SPs in CPAs; (3) examining pre-licensure students.

Selection process

Figure 1 presents the PRISMA flowchart of literature screening and the selection process. A total of 314 studies were identified from the initial search. After eliminating 121 duplicates, 193 potential studies remained. Three authors (IC, EV, and CM) independently screened titles and abstracts of the studies. We selected relevant studies after applying the inclusion criteria, and seven studies were eligible for review. Two studies identified through the reference lists of the selected articles were included. Nine studies were selected for review.

————————-Insert Figure 1————————-

Quality assessment

The *Guidelines for Critical Review (*GCR) protocol developed by the McMaster University Occupational Therapy Evidence-based Practice Research Group [67] was utilized for critical appraisal of the methodological quality and strength of evidence of selected studies. The GCR has quantitative and qualitative review guidelines though this review only applied the quantitative review guideline as all selected studies were quantitative in nature. The guideline consists of seven data extraction areas: study purpose, design, sample, outcomes, intervention, results, and conclusions and implications. The GCR protocol is available online (*https://srs-mcmaster.ca/research/evidence-based-practice-research-group/)*..

Data management, collection process, and synthesis

The authors shared the documents of the GCR instruction for assessing quality criteria, the full text of selected studies, and the results of data extraction in a private shared drive. Four authors (IC, EV, BC, and KN) reviewed each selected study and independently performed data extraction by using the GCR protocol. The first author (IC) then synthesized all the data and uploaded them to the shared drive. After an initial evaluation, there were minor discrepancies between evaluations. Discrepancies were discussed until consensus was achieved. The primary cause for discrepancies was related to the interpretation of the GCR guideline.

Results

Table 1 lists the selected studies; Table 2 shows the descriptive summary of the studies; Table 3 presents the results of a methodological critique of the selected studies using the GCR protocol.

————————-Insert Table 1————————-

————————-Insert Table 2————————-

————————-Insert Table 3————————-

Criteria 1 - Purpose and Design

All selected studies clearly stated their purpose. Most of the studies reviewed relevant background literature to justify the research however, Dewhurst et al. [27] and Stupart et al. [44] did not provide adequate background information. As the knowledge around a subject grows, study designs should become more rigorous where most variables affecting the consequence are understood and can be controlled by the researcher [67]. If there is a paucity of information about an issue, a more exploratory method is suitable; for instance, a case study or a cross-sectional design. The most rigorous experimental design is the RCT. In this regard, all the studies reviewed have appropriate study designs. Five selected studies published from 2007 to 2017 used retrospective cross-sectional design [3,27,35,44,56]. Adamson et al. [68], Schleicher et al. [51], and Wass et al. [37] utilized prospective cross-sectional designs and Yeates et al. [39] used a randomized controlled trial (RCT). The temporal distribution of the studies shows a progression from cross-sectional designs with the earliest studies to an RCT design with the most recent study.

Criteria 2—Sampling

All selected studies derived the personal information of examiners and/or students from institutional databases or asked the participants to self-declare ethnicity and/or gender. Most of the selected studies stated the ethnicity and gender of examiners and students clearly, Dewhurst et al. [27] did not describe examiners' gender and ethnicity. Five retrospective cross-sectional studies recruited participants from the previous 1 to 4 years. The number of examiners in these five studies ranged from 251 to 356 and the number of assessments ranged from 3008 to 52000. All the studies justified the sample size. Denney et al. [35] collected data on the MRCGP CSA in the UK, Dewhurst et al. [27] and McManus et al. [3] recruited participants from the MRCP (UK), and Richens et al. [56] included participants from the Intercollegiate Specialty Board examinations in the UK; these four studies were considered as representative of the population as they included all students for the specified range of time. Schleicher et al. [51], Stupart et al. [44], Wass et al. [37], and Yeates et al. [39] conducted convenience sampling at five German medical schools, a medical school in South Africa, and medical schools throughout the UK, respectively; Adamson et al. [68] used convenience sampling to recruit participants from the faculty of nursing faculty at a university in the United States.

Criteria 3—Outcome measures

Six out of nine studies used rating scales to award grades for examinees, of which Denney et al. [35], Dewhurst et al. [27], and McManus et al. [3] most clearly described the reliability and validity of their outcome measures. Schleicher et al. [51] addressed the weakness of the interrater reliability of their scale. Wass et al. [37] and Yeates et al. [39] did not describe the reliability of their rating scales. In addition, Adamson et al. [68] applied an 11-dimension rubric with good reliability and validity. Richens et al. [56] and Stupart et al. [44] did not describe the tools used to measure student performance.

Criteria 4—Implementation

Five selected studies were implemented in undergraduate medical education [3,27,35,37,39]; three studies in surgical residencies [44,51,56], and one study in nursing [68]. All studies provided prior training to examiners before the CPA or OSCE.

Seven out of nine studies were conducted in a real exam situation, where the scores assigned to students affected students' course grades or licensure. In these seven studies, the number of CPA stations ranged from 5 to 22; three studies had one examiner in each station [35,37,44], and two studies had two examiners [3,51]; Richen et al. [56] and Dewhurst et al. [27] did not mention the number of examiners per station. Two out of nine studies were in a simulated exam setting, where the scores awarded did not affect students' grades. In these two studies, Adamson et al. [68] randomly allocated 68 examiners to four student simulations; Yeates et al. [39] randomly assigned 159 examiners to two student simulation groups.

Eight of the studies describe the procedure and content of the CPA or OSCE. Dewhurst et al. [27] did not provide adequate information for the OSCE used.

Criteria 5—Results and conclusions

Denney et al. [35], Dewhurst et al. [27], Richens et al. [56], and Stupart et al. [44] demonstrated underperformance in the overall clinical performance of students from ethnic minorities compared to the ethnic majority; Wass et al. [37] found the underperformance of non-White students was restricted to communication skills. Adamson et al. [68] only included students from ethnic minorities and found no ethnic-related effects on the CPA.

Three studies reported ethnic bias: Denney et al. [35] found a significant interaction between examiners' and students' ethnicity with Black and Minority Ethnic (BME) students receiving higher grades from BME examiners than White examiners.; McManus et al. [3] reported only one non-White examiner consistently awarding higher scores to non-White students; Dewhurst et al. [27] identified two non-White examiners demonstrating a bias towards non-White students. Yeates et al. [39] found that examiners activated Asian related stereotypes, but these had no effect on examiner scorings. Although these studies demonstrated ethnic bias in clinical performance, it does not appear there is any systematic ethnic-related examiner bias due to the small effect sizes.

Regarding gender-related differences, Dewhurst et al. [27], Schleicher et al. [51], and Stupart et al. [44] found that female students performed significantly better in the overall clinical performance than male students. Two studies reported potential gender bias: Denney et al. [35] found male students receiving higher scores from female examiners compared to male examiners; Schleicher et al. [51] found male examiners scoring female students higher than male students. Two studies reported no gender-related differences or bias in CPAs [3,56].

Criteria 6—Study bias and limitations

There were four major limitations in the selected studies. First is the representativeness of the sample to a larger population. Five studies used convenience sampling [37,39,44,51,68] and were not likely representative of the population [69,70]. The second limitation was the under-reported psychometric properties of the outcome measures. One study did not state the reliability and validation of the assessment tool used in CPAs [37], one did not address the reliability of the tool [39], and two studies did not describe the assessment tools in the CPA [44,56]. This issue may lower the confidence and quality of the outcomes of interest [71]. The third limitation is the lack of descriptive details of the implementation: two studies did not describe their implementation in detail. These studies lacked information about the OCSE used and did not mention the number of examiners [27,56]. The final limitation is the use of only a single examiner per CPA station. Five studies used only one examiner per station, potentially reducing the reliability of the outcome measures [37,39,44,56,68].

Quality based on the GCR criteria

Table 4 shows the extent to which the studies met the GCR quality criteria and presents a summary of this systematic review. Adamson et al. [68], Denney et al. [35], and McManus et al. [3] met all quality criteria indicating the results of these three studies may be more trustworthy. Adamson et al. [68] did not detect any potential examiner bias. Denney et al. [35] demonstrated a significant interaction between examiners' and students' ethnicity and gender but with low effect size. McManus et al. [3] found ethnic-related bias occurring in one non-White examiner who awarded higher grades to non-White students.

————————-Insert Table 4————————-

Discussion

Based on the results of the GCR criteria, the overall quality of the studies is moderate with only three studies meeting all criteria. Although three selected studies identified examiner bias related to ethnicity, the outcomes were not consistent [27,35,58]. Dewhurst et al. [27] and Denney et al. [35] found examiner bias in only one and two examiners, respectively. Denney et al. [35] reported BME examiners favoring BME students but with a small effect size. Of these three studies, only Denney et al. [35] met all the GCR quality criteria.

Two selected studies demonstrated examiner bias related to gender, with female examiners awarding higher marks to male students [35] and male examiners awarding higher marks to female students [51], both outcomes had small effect size. Of these two studies, only Denney et al. [35] met all the GCR quality criteria. Though significant differences for CPA scores were found for ethnic and gender groups; overall, there is no consistent evidence to support ethnic or gender bias occurring in examiners assessing pre-licensure healthcare students in standardized CPAs. Nonetheless, the potential for bias in some examiners should not be ignored; frame of reference (FOR) training, using multiple examiners per station, and combined assessments are recommended for CPAs.

Predictors of Underperformance

Communication skills tend to be the most significant factor influencing clinical performance [40,43,44]. Two of the selected studies in this review identified the ethnic differences related to the communication skills among groups [27,37]. Wass and

colleagues [37] reported a small group of non-White male students using a medical model of consultation rather than a more empathetic and social style preferred by examiners. In other words, this group of students followed the textbook guideline to communicate with patients that may be less demanding of communication skills and perceived as appropriate. Additionally, female students or physicians perform better in communication and interpersonal skills with patients [52–54,59,72,73], which could be correlated to their abilities to actively listen and a greater sense of patient care values, female practitioners may also find it easier to build rapport with patients [74]. It seems that students from the ethnic majority and female students can do better in CPAs because of their greater communication or interpersonal skills. For students with poor communication or interpersonal skills, instructors can provide targeted training to improve such skills.

Suggestions for mitigating potential examiner bias

Three suggestions for mitigating potential examiner bias are (1) Frame of reference training, (2) multiple examiners, and (3) combination assessments.

Frame of reference (FOR) training

Chong and colleagues [75] suggest training to define the scoring principles and outline expected levels of student performance to reduce examiner bias in OSCEs. To our knowledge, no literature has recommended a framework for examiner training in CPAs to minimize ethnic or gender bias. The authors suggest that FOR training may be effective in eliminating bias. The aim of FOR training is to direct examiner scoring with a common standard of performance [76,77]. Evidence demonstrates that FOR training can increase the accuracy of scoring or giving feedback in CPAs [78,79].

FOR training involves creating training groups of examiners, explaining the use of the assessment tool, and iteratively practicing scoring based on standardized examples. Practicing scoring involves explaining and discussing ratings and clarifying any disagreements [76,77,80,81]:

Multiple examiners and combination assessments

The presence of multiple examiners minimizes the risk of a biased CPA [82]. The use of two or more examiners in a CPA station could increase the reliability and fairness of the results and favor the objectivity of the assessment [17,83–86]. With another examiner, an examiner's scoring could be monitored and compared making it easier to eliminate potential bias. However, using multiple examiners sometimes is not practical due to the cost of human resources; one study suggested utilizing nonmedical lay-examiners as non-medical lay examiners showed similar inter-rater reliability to trained practitioner-examiners after training [87].

Although CPAs strongly predict student performance, a combination of other assessments (i.e., essays and multiple-choice tests) have been shown to have the strongest predictive validity of student performance [88]. It is important to note that standardized CPAs have no established gold standard, thus incorporating other assessments with CPAs can help to identify discrepancies that may occur in CPAs based on ethnicity or gender.

Other potential sources of bias

SP influences

In all of the selected studies, only one study utilized both educators and SPs as examiners with the SPs scoring students' communication skills [37]. SPs can be trained to provide feedback after assessments. However, when SPs must simultaneously act while observing and scoring students' clinical performance, it may increase their mental workload and cause cognitive bias. Studies indicated negative effects for evaluation accuracy and less consistent and reliable ratings when SPs' were used that

both acted and rated performance when compared to physician- or non-medical lay- examiner ratings [89,90].However, studies have shown no or minimal SP-examiner gender bias on CPA scoring [91].

Assessment Tools

It is possible assessment tools (i.e., scales, checklists, rubrics) as opposed to the examiners themselves are responsible for the differences in scores found between ethnic or gender minorities/majorities. The use of advanced test theory methods can allow for a more objective understanding of item functioning, including performance across subgroups. Item Response Theory (IRT), Generalizability theory, and methods for identifying item bias such as Differential Item Functioning (DIF) can provide a more comprehensive understanding of the psychometric properties of the measures used for CPAs [92]. By using these methods, it will be possible to determine if sources of bias lay within the exam items rather than the examiners. Consideration must be taken to ensure the assessment items accurately and fairly reflect the assessment objectives [93].

Further, overly complex assessment tools could create bias as examiners must observe student performance and review the content of scales at the same time. Complex assessment tools might not increase measurement qualities, and the examiners' attention and efforts may be better spent on performance observation and interpretation rather than recalling how to use the assessment tools. Simplified assessment tools for OSCEs can have similar reliability and pass rates as more complex ones and using simplified assessment tools in CPAs can avoid the issue of cognitive overload for examiners [94,95].

Limitations

The primary limitation of the present review was the use of the GCR protocol. The GCR protocol appears to be seldom used in systematic reviews related to healthcare education and was initially designed for assessing the qualities of clinical intervention outcomes [66]. The GCR protocol was used for the present review as it provides clear guidelines, and its criteria were suitable for all the selected studies. Few studies investigating potential examiner bias met the criteria for high-quality studies according to GCR criteria; future studies should strive for population-representative samples, reporting of psychometric properties of assessment tools, and providing more explicit details of research implementation.

Conclusions

Based on the reviewed literature, it does not appear consistent or systematic bias based on gender or ethnicity exists. This statement should be qualified by the fact that the overall quality of the studies reviewed was moderate, with only three out of nine selected studies meeting all GCR criteria for a high-quality study. Across most of the studies, there were methodological issues including sampling, assessment tools used, and description of the CPAs studied. Further investigation of potential gender and ethnic bias using more rigorous methods is required. Further investigation is needed into the effect of FOR training and using multiple examiners per CPA station to reduce potential examiner bias.

Abbreviations

BME: Black and minority ethnic; CPA: Clinical performance assessment; CPE: Clinical performance evaluation; CPX: Clinical performance examination; CSA: Clinical skills assessment; DIF: Differential Item Functioning; FOR: Frame of reference; GCR: Guidelines for critical review; IRT: Item Response Theory; MRCGP: Membership of the Royal College of General Practitioners; MRCP (UK): Membership of the Royal Colleges of Physicians in the United Kingdom; OSCE: Objective structured clinical examination; PACES: Practical Assessment of Clinical Skills; PBA: Performance-based assessment; RCT: Randomized controlled trial; SP: Standardized patient.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
All data analyzed in this systematic review were extracted from available published articles.

Competing interests
The authors declare that they have no competing interests.

Funding
No funding was received.

Authors' contributions
Conceptualization of the systematic review: IC, EV, SK, MR. Article search: IC, EV, CM. Data extraction and analysis: IC, EV, BC, KN. Writing the manuscript: IC, EV, BC, CM. Review and editing the manuscript: IC, EV, BC, CM, SK, MR. All authors approved the final manuscript.

Acknowledgements
The authors have no acknowledgments to make.

Authors' information
[1]Faculty of Rehabilitation Medicine, University of Alberta, Edmonton T6G 2R3, Canada. [2]Department of Educational Psychology, Faculty of Education, University of Alberta, Edmonton T6G 2R3, Canada. [3]Department of Occupational Therapy, Faculty of Rehabilitation Medicine, University of Alberta, Edmonton T6G 2R3, Canada.

References

1. Wu W, Martin BC, Ni C. A Systematic Review of Competency-Based Education Effort in the Health Professions: Seeking Order Out of Chaos. Healthc Policy Reform Concepts Methodol Tools Appl. 2019;1410–36.
2. Terry R, Hing W, Orr R, Milne N. Do coursework summative assessments predict clinical performance? A systematic review. BMC Med Educ [Internet]. 2017 Feb 16 [cited 2019 Apr 22];17. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5314623/
3. McManus, Elder AT, Dacre J. Investigating possible ethnicity and sex bias in clinical examiners: an analysis of data from the MRCP(UK) PACES and nPACES examinations. BMC Med Educ. 2013 Dec;13(1):103.
4. Stegers-Jager KM. Is it them or is it us? Unravelling ethnic disparities in undergraduate clinical performance. BMC Med [Internet]. 2017 Oct 25 [cited 2019 May 7];15. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5655830/
5. Ebbert D, Connors H. STANDARDIZED PATIENT EXPERIENCES: Evaluation of Clinical Performance and Nurse Practitioner Student Satisfaction. Nurs Educ Perspect. 2004 Jan 1;25(1):12–5.
6. Imanipour M, Jalili M. Development of a comprehensive clinical performance assessment system for nursing students: A programmatic approach. Jpn J Nurs Sci. 2016;13(1):46–54.
7. Chen HC, Teherani A, O'Sullivan P. How Does a Comprehensive Clinical Performance Examination Relate to Ratings on the Medical School Student Performance Evaluation? Teach Learn Med. 2011 Jan 12;23(1):12–4.
8. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J. 1975 Feb 22;1(5955):447–51.
9. McClelland DC. Testing for competence rather than for "intelligence." Am Psychol. 1973;28(1):1–14.
10. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ. 1979;13(1):39–54.
11. Erdogan A, Dong Y, Chen X, Schmickl C, Sevilla Berrios RA, Garcia Arguello LY, et al. Development and validation of clinical performance assessment in simulated medical emergencies: an observational study. BMC Emerg Med [Internet]. 2016 Jan

15 [cited 2019 Jul 4];16. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4715281/

12. Lunenfeld E, Weinreb B, Lavi Y, Amiel GE, Friedman M. Assessment of emergency medicine: a comparison of an experimental objective structured clinical examintion with a practical examination. Med Educ. 1991 Jan;25(1):38–44.
13. Gorter S, Rethans J-J, Heijde DVD, Scherpbier A, Houben H, Vleuten CVD, et al. Reproducibility of clinical performance assessment in practice using incognito standardized patients. Med Educ. 2002;36(9):827–32.
14. Nestel D, Tabak D, Tierney T, Layat-Burn C, Robb A, Clark S, et al. Key challenges in simulated patient programs: An international comparative case study. BMC Med Educ. 2011 Sep 25;11:69.
15. Stillman P. Assessment of Clinical Skills of Residents Utilizing Standardized Patients: A Follow-up Study and Recommendations for Application. Ann Intern Med. 1991 Mar 1;114(5):393.
16. Turner TR, Scerbo MW, Gliva-McConvey GA, Wallace AM. Standardized Patient Encounters: Periodic Versus Postencounter Evaluation of Nontechnical Clinical Performance. Simul Healthc J Soc Simul Healthc. 2016 Jun;11(3):164–72.
17. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ. 2011 Dec;45(12):1181–9.
18. Ferrell BG. Clinical performance assessment using standardized patients: a primer. Fam Med. 1995 Jan;27(1):14–9.
19. Wimmers PF, Schauer GF. Validating OSCE Performance: The Impact of General Intelligence. Health Prof Educ. 2017 Dec 1;3(2):79–84.
20. Edgar S, Mercer A, Hamer P. Admission interview scores are associated with clinical performance in an undergraduate physiotherapy course: an observational study. Physiotherapy. 2014 Dec;100(4):331–5.
21. Guttormsen S, Beyeler C, Bonvin R, Feller S, Schirlo C, Schnabel K, et al. The new licencing examination for human medicine: from concept to implementation. Swiss Med Wkly [Internet]. 2013 Dec 3 [cited 2019 May 7];143(4950). Available from: https://smw.ch/en/article/doi/smw.2013.13897/
22. Sakurai H, Kanada Y, Sugiura Y, Motoya I, Wada Y, Yamada M, et al. OSCE-based Clinical Skill Education for Physical and Occupational Therapists. J Phys Ther Sci. 2014 Sep;26(9):1387–97.
23. Aranda JP, Davies ML, Jackevicius CA. Student pharmacists' performance and perceptions on an evidence-based medicine objective structured clinical examination. Curr Pharm Teach Learn. 2019 Mar;11(3):302–8.
24. Näpänkangas R, Karaharju-Suvanto T, Pyörälä E, Harila V, Ollila P, Lähdesmäki R, et al. Can the results of the OSCE predict the results of clinical assessment in dental education? Eur J Dent Educ. 2016;20(1):3–8.
25. Khan R, Payne MWC, Chahine S. Peer assessment in the objective structured clinical examination: A scoping review. Med Teach. 2017 Jul 3;39(7):745–56.
26. Zayyan M. Objective Structured Clinical Examination: The Assessment of Choice. Oman Med J. 2011 Jul;26(4):219–22.
27. Dewhurst NG, McManus C, Mollon J, Dacre JE, Vale AJ. Performance in the MRCP(UK) Examination 2003–4: analysis of pass rates of UK graduates in relation to self-declared ethnicity and gender. BMC Med. 2007 Dec;5(1):8.
28. CarlLee S, Rowat J, Suneja M. Assessing Entrustable Professional Activities Using an Orientation OSCE: Identifying the Gaps. J Grad Med Educ. 2019 Apr;11(2):214–20.
29. Franzese C. When to Cut? Using an Objective Structured Clinical Examination to Evaluate Surgical Decision-Making. The Laryngoscope. 2007;117(11):1938–42.
30. Lee CB, Madrazo L, Khan U, Thangarasa T, McConnell M, Khamisa K. A student-initiated objective structured clinical examination as a sustainable cost-effective learning experience. Med Educ Online [Internet]. 2018 Feb 26 [cited 2019 Jul 5];23(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5827782/
31. Lukas RV, Adesoye T, Smith S, Blood A, Brorson JR. Student assessment by objective structured examination in a neurology clerkship. Neurology. 2012 Aug 14;79(7):681–5.
32. Schwartz RW, Witzke DB, Donnelly MB, Stratton T, Blue AV, Sloan DA. Assessing residents' clinical performance: cumulative results of a four-year study with the Objective Structured Clinical Examination. Surgery. 1998 Aug;124(2):307–12.
33. Sloan DA, Donnelly MB, Schwartz RW, Strodel WE. The Objective Structured Clinical Examination. The new gold standard for evaluating postgraduate clinical performance. Ann Surg. 1995 Dec;222(6):735–42.

34. Wright EJ, Khosla RK, Howell L, Lee GK. Rhinoplasty Education Using a Standardized Patient Encounter. Arch Plast Surg. 2016 Sep;43(5):451–6.
35. Denney ML, Freeman A, Wakeford R. MRCGP CSA: are the examiners biased, favouring their own by sex, ethnicity, and degree source? Br J Gen Pract. 2013 Nov;63(616):e718–25.
36. Stegers-Jager KM, Steyerberg EW, Cohen-Schotanus J, Themmen APN. Ethnic disparities in undergraduate pre-clinical and clinical performance. Med Educ. 2012;46(6):575–85.
37. Wass V, Roberts C, Hoogenboom R, Jones R, Van der Vleuten C. Effect of ethnicity on performance in a final objective structured clinical examination: qualitative and quantitative study. BMJ. 2003 Apr 12;326(7393):800–3.
38. Woolf K, Cave J, Greenhalgh T, Dacre J. Ethnic stereotypes and the underachievement of UK medical students from ethnic minorities: qualitative study. The BMJ [Internet]. 2008 Aug 18 [cited 2019 May 7];337. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2517162/
39. Yeates P, Woolf K, Benbow E, Davies B, Boohan M, Eva K. A randomised trial of the influence of racial stereotype bias on examiners' scores, feedback and recollections in undergraduate clinical exams. BMC Med. 2017 Dec;15(1):179.
40. Fernandez A, Wang F, Braveman M, Finkas LK, Hauer KE. Impact of Student Ethnicity and Primary Childhood Language on Communication Skill Assessment in a Clinical Performance Examination. J Gen Intern Med. 2007 Aug;22(8):1155–60.
41. Haq I, Higham J, Morris R, Dacre J. Effect of ethnicity and gender on performance in undergraduate medical examinations. Med Educ. 2005;39(11):1126–8.
42. Woolf K, Potts HWW, McManus IC. Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis. The BMJ [Internet]. 2011 Mar 8 [cited 2019 May 6];342. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3050989/
43. Isik U, Wilschut J, Croiset G, Kusurkar RA. The role of study strategy in motivation and academic performance of ethnic minority and majority students: a structural equation model. Adv Health Sci Educ. 2018;23(5):921–35.
44. Stupart D, Goldberg P, Krige J, Khan D. Does examiner bias in undergraduate oral and clinical surgery examinations occur? 2008;98(10):3.
45. Clouten N, Homma M, Shimada R. Clinical education and cultural diversity in physical therapy: clinical performance of minority student physical therapists and the expectations of clinical instructors. Physiother Theory Pract. 2006 Jan;22(1):1–15.
46. Graham JR, West LM, Martinez J, Roemer L. The mediating role of internalized racism in the relationship between racist experiences and anxiety symptoms in a Black American sample. Cultur Divers Ethnic Minor Psychol. 2016;22(3):369–76.
47. Hayes SC, Strosahl KD, Wilson KG. Acceptance and commitment therapy: An experiential approach to behavior change. New York, NY, US: Guilford Press; 1999. xvi, 304. (Acceptance and commitment therapy: An experiential approach to behavior change).
48. Herbert JD, Cardaciotto L. A MINDFULNESS AND ACCEPTANCE-BASED PERSPECTIVE ON SOCIAL ANXIETY DISORDER. 2005;24.
49. Rapee RM, Heimberg RG. A cognitive-behavioral model of anxiety in social phobia. Behav Res Ther. 1997 Aug 1;35(8):741–56.
50. Jacques L, Kaljo K, Treat R, Davis J, Farez R, Lund M. Intersecting gender, evaluations, and examinations: Averting gender bias in an obstetrics and gynecology clerkship in the United States. Educ Health. 2016;29(1):25.
51. Schleicher, Leitner K, Juenger J, Moeltner A, Ruesseler M, Bender B, et al. Examiner effect on the objective structured clinical exam—a study at five medical schools. BMC Med Educ. 2017 Dec;17(1):71.
52. Blanch DC, Hall JA, Roter DL, Frankel RM. Medical student gender and issues of confidence. Patient Educ Couns. 2008 Sep 1;72(3):374–81.
53. Cuddy MM, Swygert KA, Swanson DB, Jobe AC. A Multilevel Analysis of Examinee Gender, Standardized Patient Gender, and United States Medical Licensing Examination Step 2 Clinical Skills Communication and Interpersonal Skills Scores. Acad Med [Internet]. 2011 Oct 1 [cited 2019 May 1];86(10). Available from: insights.ovid.com

54. Riese A, Rappaport L, Alverson B, Park S, Rockney RM. Clinical Performance Evaluations of Third-Year Medical Students and Association With Student and Evaluator Gender. Acad Med J Assoc Am Med Coll. 2017;92(6):835–40.
55. Woolf K, Haq I, McManus IC, Higham J, Dacre J. Exploring the underperformance of male and minority ethnic medical students in first year clinical examinations. Adv Health Sci Educ. 2008 Dec 1;13(5):607–16.
56. Richens D, Graham TR, James J, Till H, Turner PG, Featherstone C. Racial and Gender Influences on Pass Rates for the UK and Ireland Specialty Board Examinations. J Surg Educ. 2016 Jan;73(1):143–50.
57. Austin EJ, Evans P, Magnus B, O'Hanlon K. A preliminary study of empathy, emotional intelligence and examination performance in MBChB students. Med Educ. 2007;41(7):684–9.
58. McManus, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. BMC Med Educ. 2006 Aug 18;6:42.
59. Wiskin CMD, Allan TF, Skelton JR. Gender as a variable in the assessment of final year degree-level communication skills. Med Educ. 2004;38(2):129–37.
60. Niehaus DJ, Jordaan E, Koen L, Mashile M, Mall S. Applicability and fairness of the oral examination in undergraduate psychiatry training in South Africa. Afr J Psychiatry. 2012 Mar;15(2):119–23.
61. Ong TQ, Kopp JP, Jones AT, Malangoni MA. Is There Gender Bias on the American Board of Surgery General Surgery Certifying Examination? J Surg Res. 2019 May;237:131–5.
62. Muthusami A, Mohsina S, Sureshkumar S, Anandhi A, Elamurugan TP, Srinivasan K, et al. Efficacy and Feasibility of Objective Structured Clinical Examination in the Internal Assessment for Surgery Postgraduates. J Surg Educ. 2017 Jun;74(3):398–405.
63. Plakiotis C. Objective Structured Clinical Examination (OSCE) in Psychiatry Education: A Review of Its Role in Competency-Based Assessment. In: Vlamos P, editor. GeNeDis 2016. Springer International Publishing; 2017. p. 159–80. (Advances in Experimental Medicine and Biology).
64. Chesser A, Cameron H, Evans P, Cleland J, Boursicot K, Mires G. Sources of variation in performance on a shared OSCE station across four UK medical schools. Med Educ. 2009;43(6):526–32.
65. Esmail A, Roberts C. Academic performance of ethnic minority candidates and discrimination in the MRCGP examinations between 2010 and 2012: analysis of data. BMJ. 2013 Sep 26;347:f5662.
66. PRISMA-P Group, Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev [Internet]. 2015 Dec [cited 2019 Nov 4];4(1). Available from: https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/2046–4053–4–1
67. Law M, Stewart D, Pollock N, Letts L, Bosch J, Westmorland M. Critical Review Form—Quantitative Studies. 1998;
68. Letts L, Wilkins S, Law M, Stewart D, Bosch J, Westmorland M. Guidelines for Critical Review Form: Qualitative Studies (Version 2.0). Canchild. 2007 Jan 1;
69. Adamson K. Rater Bias in Simulation Performance Assessment: Examining the Effect of Participant Race/Ethnicity. Nurs Educ Perspect N Y. 2016 Apr;37(2):78–82.
70. Bornstein MH, Jager J, Putnick DL. Sampling in Developmental Science: Situations, Shortcomings, Solutions, and Standards. Dev Rev DR. 2013 Dec;33(4):357–70.
71. Setia MS. Methodology Series Module 5: Sampling Strategies. Indian J Dermatol. 2016;61(5):505–9.
72. Souza AC de, Alexandre NMC, Guirardello E de B, Souza AC de, Alexandre NMC, Guirardello E de B. Psychometric properties in instruments evaluation of reliability and validity. Epidemiol E Serviços Saúde. 2017 Sep;26(3):649–59.
73. Nicolai J, Demmel R. The impact of gender stereotypes on the evaluation of general practitioners' communication skills: an experimental study using transcripts of physician-patient encounters. Patient Educ Couns. 2007 Dec;69(1–3):200–5.
74. Rutala PJ, Witzke DB, Leko EO, Fulginiti JV. The influences of student and standardized patient genders on scoring in an objective structured clinical examination. Acad Med. 1991;66(9, Suppl):28–30.
75. Zaharias G, Piterman L, Liddell M. Doctors and Patients: Gender Interaction in the Consultation: Acad Med. 2004 Feb;79(2):148–55.

76. Chong L, Taylor S, Haywood M, Adelstein B-A, Shulruf B. Examiner seniority and experience are associated with bias when scoring communication, but not examination, skills in objective structured clinical examinations in Australia. J Educ Eval Health Prof. 2018 Jul 18;15:17.
77. Newman LR, Brodsky D, Jones RN, Schwartzstein RM, Atkins KM, Roberts DH. Frame-of-Reference Training: Establishing Reliable Assessment of Teaching Effectiveness. J Contin Educ Health Prof. 2016;36(3):206–10.
78. Schleicher, Day DV, Mayes BT, Riggio RE. A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. J Appl Psychol. 2002 Aug;87(4):735–46.
79. Gorman A, Rentsch JR. Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. J Appl Psychol. 2009;94(5):1336–44.
80. Woehr DJ, Huffcutt AI. Rater training for performance appraisal: A quantitative review. J Occup Organ Psychol. 1994;67(3):189–205.
81. Bernardin J, Buckley R. Strategies in rater training. Acad Manage Rev. 1981;6(2):205–12.
82. Gardner AK, Russo MA, Jabbour II, Kosemund M, Scott DJ. Frame-of-reference training for simulation-based intraoperative communication assessment. Am J Surg. 2016 Sep 1;212(3):548–551.e2.
83. Bartfay WJ, Rombough R, Howse E, LeBlanc R. The OSCE approach in nursing education: Objective structured clinical examinations can be effective vehicles for nursing education and practice by promoting the mastery of clinical skills and decision-making in controlled and safe learning environments. Can Nurse. 2004 Mar;100(3):18–223.
84. Bagnasco A, Tolotti A, Pagnucci N, Torre G, Timmins F, Aleo G, et al. How to maintain equity and objectivity in assessing the communication skills in a large group of student nurses during a long examination session, using the Objective Structured Clinical Examination (OSCE). Nurse Educ Today. 2016 Mar 1;38:54–60.
85. Dickter DN, Stielstra S, Lineberry M. Interrater Reliability of Standardized Actors Versus Nonactors in a Simulation Based Assessment of Interprofessional Collaboration. Simul Healthc J Soc Simul Healthc. 2015 Aug;10(4):249–55.
86. Harasym PH, Woloschuk W, Cunning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. Adv Health Sci Educ. 2008 Dec 1;13(5):617–32.
87. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med. 2003;15(4):270–92.
88. Berger AJ, Gillespie CC, Tewksbury LR, Overstreet IM, Tsai MC, Kalet AL, et al. Assessment of medical student clinical reasoning by “lay” vs physician raters: inter-rater reliability using a scoring guide in a multidisciplinary objective structured clinical examination. Am J Surg. 2012 Jan 1;203(1):81–6.
89. Wilkinson TJ, Frampton CM. Comprehensive undergraduate medical assessments improve prediction of clinical performance. Med Educ. 2004;38(10):1111–6.
90. Newlin-Canzone E, Scerbo M, Gliva-McConvey G, Wallace A. The Cognitive Demands of Standardized Patients: Understanding Limitations in Attention and Working Memory With the Decoding of Nonverbal Behavior During Improvisations. Simul Healthc J Soc Simul Healthc. 2013 Aug;8(4):207–14.
91. Zanetti M, Keller L, Mazor K, Carlin M, Alper E, Hatem D, et al. Using standardized patients to assess professionalism: a generalizability study. Teach Learn Med. 2010 Oct;22(4):274–9.
92. Colliver JA, Vu NV, Marcy ML, Travis TA, Robbs RS. Effects of examinee gender, standardized-patient gender, and their interaction on standardized patients’ ratings of examinees’ interpersonal and communication skills. Acad Med J Assoc Am Med Coll. 1993 Feb;68(2):153–7.
93. Baig LA, Violato C. Temporal stability of objective structured clinical exams: a longitudinal study employing item response theory. BMC Med Educ. 2012 Dec 7;12:121.
94. Clauser B, Ross L, Fletcher E, Klass D, Finkbiner R, King A. Differential item functioning in checklist items from a standardized-patient-based examination. Acad Med [Internet]. 1994 Oct [cited 2019 Aug 15];69(10). Available from: insights.ovid.com

95. Hurley KF, Giffin NA, Stewart SA, Bullock GB. Probing the effect of OSCE checklist length on inter-observer reliability and observer accuracy. Med Educ Online [Internet]. 2015 Oct 20 [cited 2019 Aug 19];20. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4613902/
96. Sandilands DD, Gotzmann A, Roy M, Zumbo BD, De Champlain A. Weighting checklist items and station components on a large-scale OSCE: is it worth the effort? Med Teach. 2014 Jul;36(7):585–90.

Tables

Table 1. Reviewed studies.

Authors	Title	Year	Country
Adamson [68]	Rater bias in simulation performance assessment: examining the effect of participant race/ethnicity.	2016	US
Denney, Freeman and Wakeford [35]	MRCGP CSA: are the examiners biased, favoring their own by sex, ethnicity, and degree source?	2013	UK
Dewhurst, McManus, Mollon, Dacre and Vale [27]	Performance in the MRCP (UK) Examination 2003–4: analysis of pass rates of UK graduates in relation to self-declared ethnicity and gender.	2007	UK
McManus, Elder and Dacre [3]	Investigating possible ethnicity and sex bias in clinical examiners: an analysis of data from the MRCP (UK) PACES and nPACES examinations.	2013	UK
Richens et al. [56]	Racial and gender influences on pass rates for the UK and Ireland specialty board examinations.	2016	UK
Schleicher et al. (51)	Examiner effect on the objective structured clinical exam – a study at five medical schools.	2017	Germany
Stupart, Goldberg, Krige and Kahn (44)	Does examiner bias in undergraduate oral and clinical surgery examinations occur?	2008	South Africa
Wass, Roberts, Hoogenboom, Jones and Van der Vleuten (37)	Effect of ethnicity on performance in a final objective structured clinical examination: qualitative and quantitative study.	2003	UK
Yeates et al. (39)	A randomised trial of the influence of racial stereotype bias on examiners' scores, feedback and recollections in undergraduate clinical exams.	2017	UK

Note: CSA = Clinical Skills Assessment; MRCGP = Membership of the Royal College of General Practitioners; MRCP = Membership of the Royal Colleges of Physicians; nPACES = new Practical Assessment of Clinical Examination Skills; PACES = Practical Assessment of Clinical Examination Skills; UK = United Kingdom; US = United States.

Descriptive summary of the reviewed studies.

	Purpose	Design	Implementation	Results	Conclusions
on	· To explore whether the scores given by examiners to simulated students were affected by students' ethnicity, while evaluating clinical judgment.	· Cross-sectional (prospective) · N of examiners = 68 · N of assessments = 68 (17 rating results in each scenario) · Ethnicity of students: 1 East Indian, 1 Latino, 1 Filipina and 1 African American · Discipline: nursing · Assessment type: a CPA · Tool: LCJR with 11 dimensions	· The examiners viewed a 22-minute LCJR training video. · Examiners randomly viewed one out of four 14-minute simulation scenario and rated using the LCJR.	· No significant differences between the scores in the four scenarios.	· The evaluation of clinical performance using the LCJR for nursing students was not affected by students' ethnic background.
et	· To determine whether the gender and ethnicity of students affected examiners' ratings.	· Cross-sectional (retrospective) · N of examiners = 251 · N of assessments = 52,000 (each student encountered 13 cases, in total 4000 students) · Ethnicity of students: 1586/ 2414 (White/ BME) · Gender of students: 2044/ 1956 (Female/ Male) · Discipline: medicine · Assessment type: an OSCE · Tool: a 4-point scale (Clear Fail, Fail, Pass, Clear Pass)	· Prior training of the examiners and simulated patients took place before the examination. · The OSCE consisted of 13 stations. · The examiners rated students on their own independent of other examiners.	· BME students received significant lower scores compared to the White students. · BME students received higher scores from BME examiners than white examiners. · Male students received higher scores from female examiners compared to male examiners. · The effect sizes were very small.	· Although there were significant differences in examiner marking behaviour, the student subgroups did not arise from the scoring behaviour of specific examiner subgroups. · Systematic bias by examiner subgroups did not explain any substantial differential student subgroup performance.
st 7)	· To investigate the effects of gender and ethnicity on medical students' pass rates.	· Cross-sectional (retrospective) · N of assessments = 3008 (in total 2353 students) · Ethnicity of students: 46/ 453/ 203/ 43/ 53/ 1704/ 26/ 480 (Afro-Caribbean/ Asian sub-continent/ Far East/ Middle Eastern/ Mixed, White/ Other/ Unknown)	· 14 assessments were made for two communication stations and three clinical skills stations. · Ten examiners assessed one student at these five stations.	· The Caucasian group had a significantly higher pass rate than all other groups combined. · No significant differences for scores between the non-White groups. · Female students performed better than male students. · Non-White males did less well than non-	· Two non-White examiners showed bias. · The effects of ethnicity were very small and restricted to communication skills. · Although female students performed better than males, the reason is likely multi-factorial. There is no evidence of gender-related bias.

		· Gender of students: 1467/ 1541 (Female/ Male) · Discipline: medicine · Assessment type: an OSCE · Tool: a 4-point scale (Clear Fail, Fail, Pass, Clear Pass)		White female or White male students. · No interaction of clinical skills with ethnicity or gender. · Significant interaction of communication skills with ethnicity.	
us)	· To report an innovative method to assess ethnic and gender bias in the PACES and nPACES examinations.	· Cross-sectional (retrospective) · N of examiners = 1790 · N of students = 17,442 · Ethnicity of students: 7028/ 7228 (non-White/ White) · Gender of students: 7936/ 9506 (Female/ Male) · Discipline: medicine · Assessment type: OSCEs · Tool: a 4-point scale (Clear Fail, Fail, Pass, Clear Pass) for PACES, and a 3-point scale (Unsatisfactory, Borderline, Satisfactory) for nPACES	· Five stations in PACES, each station gave a score on a four-point scale. · Five stations in nPACES assessing 4 to 7 skills, each skill was assessed on a three-point scale. · Two examiners in each station of both exams.	· There was no apparent favoring of male and favoring female students. were similar. · One non-White examiner showed ethnic bias consistently, giving higher scores to non-White students.	· There was no overall gender bias. · The new method worked well to explore potential examiner bias in examinations with more than one examiner per station.
s et	· To determine whether gender, ethnic origin, training status and first language affected pass rates. · To explore whether students had similar results either marked by a computer and by examiners.	· Cross-sectional (retrospective) · N of examiners = 3567 (section 2) · N of assessments = 5035 (section 2) · Ethnicity of students: 36%/ 3%/ 2%/ 27%/ 11%/ 22% (Asian/ Black/ Mixed/ White/ Other/ Prefer not to say) · Gender of students: 12%/ 87%/ 2% (Female/ Male/ Prefer not to say) · Discipline: surgery · Assessment type: an MCQ exam, and an OSCE · Tool: not known	· Nine specialties were included in the examination. · Each specialty consists of two sections. Section 1 a multiple-choice exam was marked by computer; section 2 a patient-based clinical exam was marked by trained examiners.	· Caucasian students got higher marks in both computer-based and face-to-face examinations however no significant difference was found. · Females performed better in face-to-face exams but not significantly.	· There were significant differences in for ethnicity and no potential examiner bias was found.
her	· To determine if	· Cross-sectional	· Part A (performance and	· Female students	· Gender bias was

21)	exam scoring was biased by the students and examiners' gender.	(prospective) · N of examiners = 5 local examiner + 1 reference examiner (1 local examiner in each medical school, there are five medical schools recruited) · N of assessments = 540 · Gender of students: 321/ 219 (Female/Male) · Discipline: surgery · Assessment type: an OSCE · Tools: 5-step-Likert-scale	knowledge) and B (communication and interaction) checklists were included in the OSCE. · Students in each school were scored by a reference examiner and a local examiner.	had overall higher scores compared to males. · Male examiners scored female students higher than male. · The effect size of gender bias was weak.	detected but with a small effect size.
et	· To explore whether any bias according to gender, ethnicity and language occurred on a CPA.	· Cross-sectional (retrospective) · N of students = 604 · Ethnicity of students: 170/ 99/ 102/ 233 (African/ coloured/ Indian/ White) · Gender of students: 369/ 235 (Female/ Male) · Discipline: surgery · Assessment type: an OSCE, a long case clinical examination, and an unstructured oral examination · Tools: not known	· The OSCE contained 20 stations, with each station having a different examiner. · A long case clinical examination occurred consisting of history taking and assessment. · In the unstructured oral examination, two examiners evaluated students on a range of surgical topics.	· Significant differences were found in the scores between population groups in all the exams. Caucasian students scored the highest, Black students scored the lowest. · Female students were awarded significantly higher marks than male on the OSCE exams.	· There was no evidence of any examiner bias in clinical or oral examinations.
t	· To investigate whether ethnic bias affects the grades of students from ethnic minorities in a final-year OSCE.	· Cross-sectional (prospective) · N of students = 175 · Ethnicity of students: 50/31/1/6 (White/ South Asian/ Afro-Caribbean/ Other) · Discipline: medicine · Assessment type: an OSCE · Tools: a 5-point scale	· A final-year OSCE conducted in a medical school contained two stations for history taking, nine for clinical examination, six for communication skills, and five for practical skills. · An independent examiner scored students at each station. SPs scored students communication skills. · Videos were recorded for examiners and SPs discussions to explore if there was any ethnic discrimination.	· In the communication stations, the grades of students from the ethnic minorities were significantly lower than Caucasian students. · No examples of ethnic discrimination occurred in the video recordings.	· There was no evidence of ethnic bias occurring.
et	· To explore whether students' scores or feedback show	· RCT · N of examiners = 159	· Examiners were randomly assigned to watch performances from White and Asian students that were either	· Examiners responded to Asian-stereotypical words faster than neutral words, suggesting	· Examiner bias did not appear to explain the differential attainment of Asian

influence of ethnic bias. · To explore whether examiners unconsciously activate stereotypes when judging Asian students' performance.	· Ethnicity of simulated students: 2 Asians, 2 White · Discipline: medicine · Assessment type: an OSCE · Tools: a 7-point scale (Fail 1-2, Borderline 3, Pass 4, Good 5, Excellent 6-7)	consistent or inconsistent with a previously described stereotype of Asian students' performance. · Examiner groups scores and comments they gave Caucasian and Asian students were compared.	Asian stereotypes were activated in examiners' minds.	students in UK medical schools.

BS = American Board of Surgery; BME = Black and minority ethic; CPA = Clinical Performance Assessment; CSA = Clinical Skills Assessm
ısater Clinical Judgment Rubric; MCQ = Multiple Choice Question; MRCGP = Membership of the Royal College of General Practitioners; r
Practical Assessment of Clinical Skills; OSCE = Objective Structured Clinical Examination; PACES = Practical Assessment of Clinical Skil
domized Control Trial; SP = Standardized Patient.

Results of methodological critique of the reviewed studies using the GCR protocol.

ıality	Adamson (69)[a]	Denney et al. (35)	Dewhurst et al. (27)	McManus et al. (3)	Richens et al. (56)	Schleicher et al. (51)[b]	Stupart et al. (44)	Wass et al. (37)[b]	Yeates et al. (39)[b]
ɔurpose:									
d ɣ?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ʻant ture ɯed?	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
:									
	Cross-sectional (prospective)	Cross-sectional (retrospective)	Cross-sectional (retrospective)	Cross-sectional (retrospective)	Cross-sectional (retrospective)	Cross-sectional (prospective)	Cross-sectional (retrospective)	Cross-sectional (prospective)	RCT
ɔpriate?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ɛ:									
led?	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
ied?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ıe ɾes:									
ble?	Yes	Yes	Yes	Yes	Not addressed	No	Not addressed	Not addressed	Not addressed
?	Yes	Yes	Yes	Yes	Not addressed	Yes	Not addressed	Not addressed	Yes
ıentation:									
ɾibed in l?	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes
ɜ:									
rted with ɜtical ficance?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
sis ɔpriate?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ɾtance ted?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
outs ted?	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Yes
lusions ɛations									
lusion ɔpriate?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ll the ɣ ia?	Yes	Yes	Unmet two criteria	Yes	Unmet two criteria	Unmet one criterion	Unmet two criteria	Unmet one criterion	Unmet one criterion

ɛ: RCT = Randomized Controlled Trial.

1. Adamson (2016) only investigated gender bias.
2. Schleicher et al. (2017), Wass et al. (2003), and Yeates et al. (2018) only investigated ethnic bias.

ɔle 4. Summary of the systematic review.

thor ate)	Students from the ethnic majority performed significantly better than the ethnic minorities	Ethnic bias found	Female students performed significantly better than males	Gender bias found	Met all the quality criteria (Table 3)
amson 9)	✗	✗	N/A	N/A	Yes
nney et (35)	✓	✓[a]	✓	✓[a]	Yes
whurst al. 007)	✓	✓[b]	✓	✗	Unmet two criteria
cManus al. 013)	✓	✓[c]	✗	✗	Yes
chens et 015)	✓	✗	✗	✗	Unmet two criteria
hleicher al. 017)	N/A	N/A	✓	✓[d]	Unmet one criterion
upart et 008)	✓	✗	✓	✗	Unmet two criteria
ass et 003)	✓[e]	✗	N/A	N/A	Unmet one criterion
ates et 017)	✗[f]	✗	N/A	N/A	Unmet one criterion

1. Black and minority ethnic (BME) examiners gave higher scores to BME students; female examiners gave higher scores to male students. Both outcomes had small effect size.
2. Only two non-White examiners had prejudice to non-White students but with small effect size.
3. Only a non-White examiner showed consistent ethnic bias who awarded higher scores to non-White students.
4. Male examiners significantly gave higher scores to female students but with small effect size.
5. The ethnic differences were only related to communication skills.
6. Asian stereotypes were cognitively activated in examiners' but did not affect the scorings.

Figures

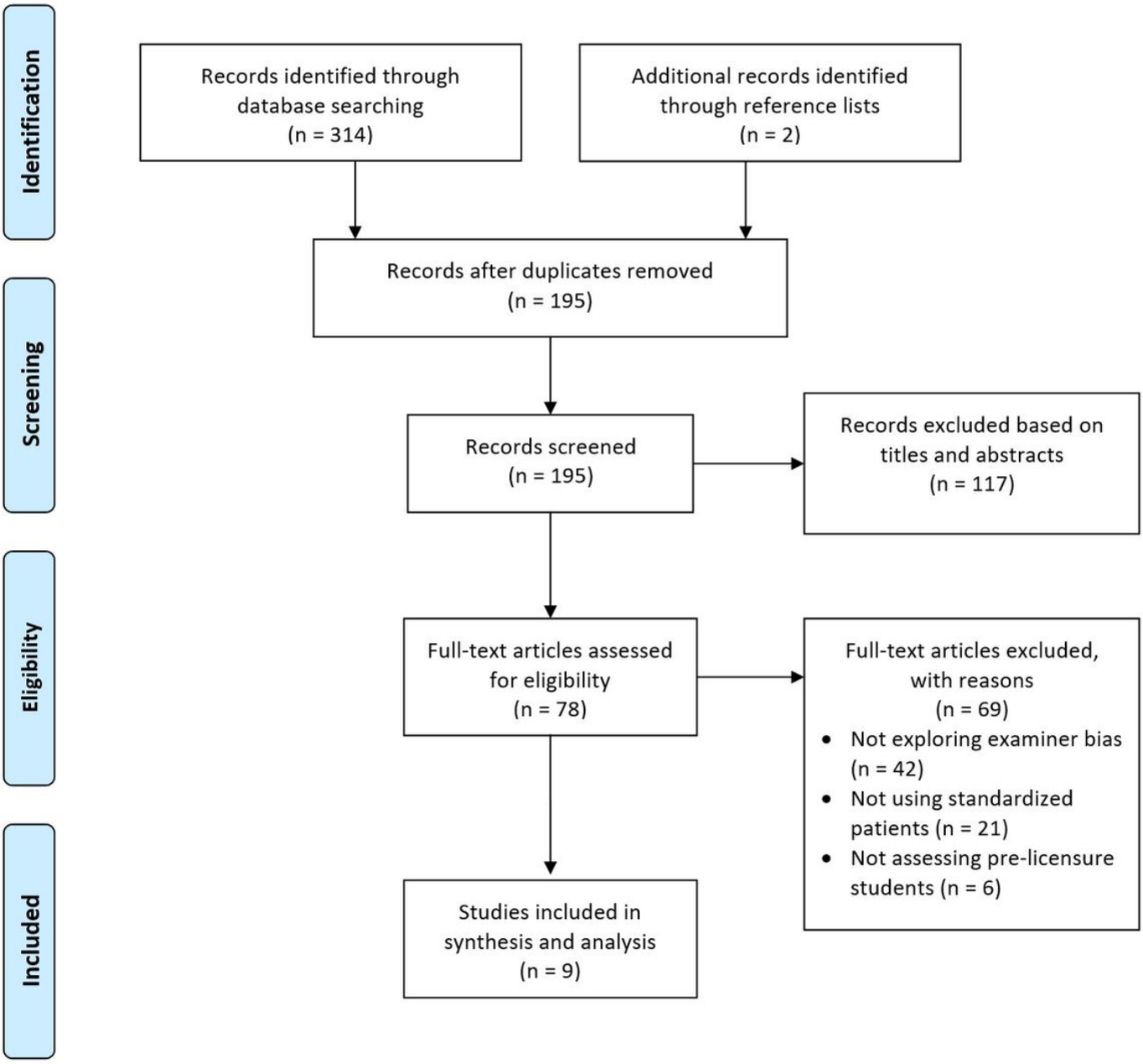

Figure 1

PRISMA flow diagram