BMC Medical Education

RESEARCH

Open Access

Impact of gender on self-assessment accuracy among fourth-year French medical students on faculty's online Objective Structured Clinical Examinations

Sylvain Bodard[1,2,3], Donia Bouzid[4], Valentine Marie Ferré[5], Claire Carette[6], Joelle Kivits[7], Yann Nguyen[8,9], Michael Thy[10,11], Ugo Marchèse[12], Bénédicte Oulès[13], Lina Khider[14], Christian de Tymowski[15,16], Nelly Burnichon[17,18], Tristan Mirault[14], Albert Faye[19] and Laura I. Levi[20*]

Abstract

Background Historically, women have been shown to underestimate their abilities, while men often assess themselves more accurately or overestimate. This study aims to determine self-assessment accuracy during online Objective Structured Clinical Examinations (OSCEs) according to gender.

Methods A prospective study was conducted among fourth-year medical students at Paris Cité University during faculty training OSCEs, utilizing Zoom® software for remote participation. Students and evaluators assessed performances using 5-point Likert scales for medical knowledge, interpersonal skills, and overall performance. Additionally, students predicted their grade out of twenty. The assessment covered three independent stations.

Results This study included 259 medical students (177 women, 81 men, one non-binary (excluded from further analyses)) evaluated by 130 physicians. Evaluator scores did not differ according to students' gender (total score out of 20: men: 10.25 ± 3.45, women: 10.23 ± 3.44 $p = 0.817$) nor students' self-assessments (total score out of 20: men: 11.22 ± 3.02, women: 11.00 ± 3.03; $p = 0.466$) whatever the domains and stations (all $p > 0.05$). The difference (delta) between self-assessment and evaluator scores for medical knowledge (men: 0.73 ± 1.00, women: 0.64 ± 1.02; $p = 0.296$), interpersonal skills (men: 1.02 ± 1.06, women: 0.93 ± 1.09; $p = 0.296$), and total score (men: 0.98 ± 3.41, women: 0.68 ± 3.42; $p = 0.296$) showed no gender differences.

Further analysis categorized students based on their self-assessment accuracy, revealing that both men and women displayed a high ratio of accurate self-assessments (78.1% for overall performance across all stations), with minimal overestimation observed in both genders (20.9% for overall performance across all stations). Instances of overestimation or underestimation were rare and not consistent over the 3 stations, indicating that such misjudgments are likely situational rather than inherent traits.

Discussion This study reveals similar self-assessment accuracy according to gender in online training OSCEs suggesting a shift towards gender-equitable self-perceptions among medical students compared to previous studies.

*Correspondence:
Laura I. Levi
laura.levi@aphp.fr
Full list of author information is available at the end of the article

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Research remains necessary to corroborate these results and explore the underlying factors contributing to this shift in self-perception.

Keywords Objective structured clinical examinations, Gender difference, Self-perception, Self-assessment, Overestimation, Underestimation

Background

Effective integration of clinical skills in medical education is crucial for training competent physicians who deliver quality care. Ronald M. Harden introduced Objective Structured Clinical Examinations (OSCEs) in 1975 to assess medical students' clinical abilities through timed stations that simulate specific clinical situations [1, 2]. The strength of OSCEs lies in their rigorous evaluation of medical knowledge and clinical competence, including decision-making, communication, and technical skills [1–4]. OSCEs have been introduced for all French medical students at the end of their sixth and last year of undergraduate, as part of the national classifying exam since 2024. To ensure adequate preparation, students undergo regular training throughout their studies. At Paris-Cité University, students participate in an annual faculty-led evaluative OSCE between their third and sixth years of study. Additionally, they complete an evaluative OSCE at the end of each internship to further refine their clinical skills and readiness.

Self-assessment is essential in pedagogy, fostering reflective, accountable learners who actively engage in their professional development. This process fosters students to engage with their learning, identify their strengths and weaknesses, and proactively shape their clinical skills. Effective self-assessment hinges on a thorough understanding of professional standards and expected competencies and the ability to objectively evaluate one's performance. Overestimation or excessive self-confidence can result in diagnostic errors and suboptimal patient care, while underestimation and increased anxiety can significantly impair student abilities and performance [5–7]. Therefore, medical students must critically evaluate their abilities because their clinical competence can directly impact on patients care [8, 9].

The literature consistently points to gender differences in self-assessment. Women often underestimate their abilities despite performing better, whereas men generally assess themselves more accurately or tend to overestimate [7, 8, 10–13]. Women also show higher anxiety about their skills, workload, and evaluations and are more concerned about their communication skills [7, 8, 14–21]. Among the reasons suggested for women's underestimations are heightened evaluation anxiety and the weight of social expectations that discourage them from appearing overconfident to avoid being perceived as pretentious [8, 14–22]. These dynamics have significant consequences for women, including a higher risk of burnout [14] due to the pressures of self-criticism and limited access to positions of responsibility or high income due to diminished self-confidence [8, 17, 23, 24]. However, with society moving toward more open, less dogmatic, and inclusive discussions of gender, it is critical to question whether these discrepancies in self-assessment persist among medical students in 2024. This study aims to assess self-assessment accuracy according to gender among Paris Cité University fourth-year medical students during faculty training online OSCEs.

Methods

Study characteristics and participants

This prospective study, approved by the Paris Cité University ethics committee (CER U-Paris 2023–10-LEVI), was conducted with fourth-year medical students as a curriculum component during their faculty training online OSCEs, part of their General Medicine class. These OSCEs are optional and students are asked a few weeks before the training if they intend to participate.

The training was conducted the 19th of January 2024 using Zoom software® (Zoom Video Communications, San Jose, CA), allowing both students and evaluators to connect from their chosen locations in a remote OSCE [25]. Each student was evaluated across three stations primarily focused on semiology (as online OSCEs do not allow for the assessment of manual skills). The stations covered the following topics: ventricular tachycardia, diarrhea, and asthma (Figures S1, S2 & S3 and Tables S1, S2 & S3). Evaluators received the station 2 days prior to the evaluation, with an online session organized to clarify the overall organization of the training and provide detailed instructions for each station. Each station involved two academic hospital physicians per station, one acting as a standardized patient, both jointly assessing the student. The data collected included the age and declared gender (female, male, or other) of each student and evaluator at the time of the OSCEs.

Two days before the test, all registered students and evaluators received an email inviting them to

participate in a study that aimed to assess determinants impacting self-assessment, without specific mention of gender. Immediately after each station, students were asked to complete an online survey for self-assessment, while evaluators graded together the student on the faculty website, Theia (THEIA, Paris, France). Subsequently, evaluators had 2 min to assess the student's performance before the student moved to the next station and continue the process. Participants were only informed about the study's primary objective at its conclusion to avoid bias. They were then invited to contact the principal investigator if they wished to withdraw from the study (Fig. 1).

Evaluation and self-assessment

The evaluations and self-assessments were based on four grading scales [12, 26]. Three of these scales were 5-point Likert scales (1 representing insufficient performance, 5 representing outstanding performance) used to evaluate medical knowledge (Table S1), interpersonal skills (Table S2), and overall performance (Table S1) separately. The fourth scale was a grade out of twenty. For the evaluator, this grade represented the final score obtained in the station, determined using a specific grading grid predetermined by the faculty (consisting of 10 to 12 items combined with 3–5 Likert scales, leading to a 20-point scale with two-third of the grade on

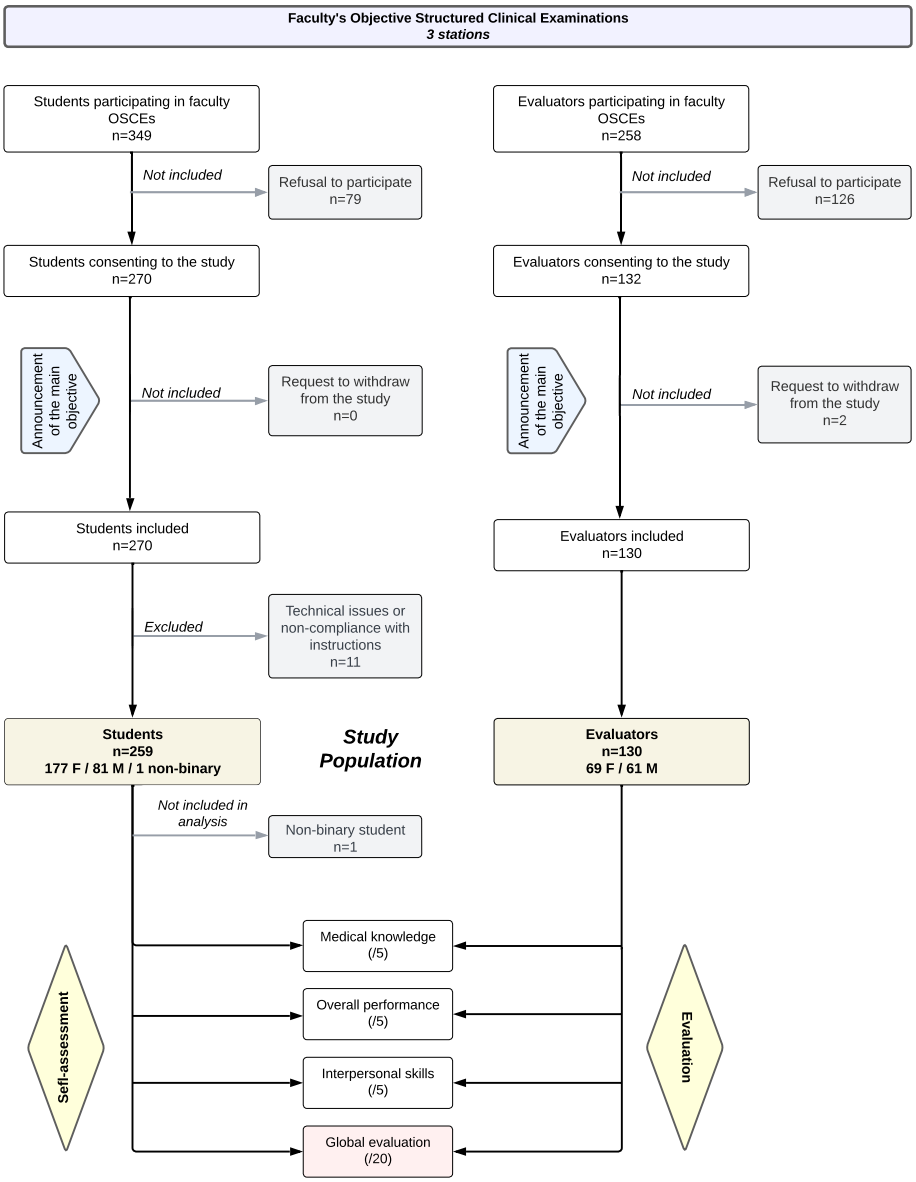

Fig. 1 Study flowchart

clinical skills and one-third on attitude and communication skills, Tables S3, S4, S5). For the student, it was a prediction of this grade that he had to guess without knowing the content of the grid. Complete survey filled in by the students is available in table S6.

Data analysis and statistics

Anonymized results were entered into an Excel spreadsheet accessible only by the principal investigator with a protected password. For each student, the delta was defined as the difference between the score the student self-assigned and the score from evaluators. It was positive in case of overestimation and negative in case of underestimation. Students were defined as overestimating and underestimating themselves in case this delta was equal or was greater than 2 (in absolute value) for Likert scale values and was strictly greater than 3 for the total score (Sullivan et Artino, J Grad Med Educ. 2013). Data analyses were performed using Prism 9 software (GraphPad Software, Boston, MA). Results are presented as mean ± standard deviation or number (percentage). Multiple t-tests were conducted to compare evaluations between female and male students, based either on raw scores or deltas, p-value were corrected for multiple comparisons. Two way ANOVA tests were used to compare proportions. Fishers' exact test was used to compare the proportion of students recurrently overestimating themselves. Redundant overestimation was defined as at least 2 stations in which a particular student overestimates himself. No redundant overestimation was defined as 0 or 1 station with overestimation.

Results

Identical performances between men and women, according to the evaluators

Among the 850 students registered in the faculty class, 349 (41%) participated in the optional online OSCE assessment. Of these, 270 (77%) agreed to participate in the study and were included. None withdrew their consent after being informed about the study objectives. Eleven students were excluded due to technical issues or non-compliance with instructions (Fig. 1).

The study participants self-declared as: men ($n=81$, 31.4%), women ($n=177$, 68.6%), or non-binary ($n=1$, 0.4%). The non-binary participant was not included in the gender comparisons. Among the 258 students included, mean age was 22.3 ± 2.5 years, with 22.2 ± 2.4 years for men versus (vs.) 22.5 ± 2.5 years for women. One hundred thirty evaluators (69 women 53.5%, 60 men 46.5%) of 42.8 ± 11.0 years old participated in the study after 2 withdraw their consent once knowing its main objective. Table S7 summarizes the evaluator functions.

The total score by the evaluators did not differ between women and men: 10.25 ± 3.45 vs.10.23 ± 3.44 out of 20 ($p=0.817$). Total scores for station 1, 2 or 3 were similar ($p=0.178$, $p=0.231$, $p=0.917$ respectively). In details, medical knowledge, interpersonal skills, and overall performance did not differ between women and men: 1.91 ± 0.88 vs. 1.87 ± 0.97 out of 5, 2.17 ± 0.84 vs. 2.20 ± 0.85 out of 5, and 1.97 ± 0.86 vs. 1.96 ± 0.95 out of 5, respectively ($p=0.734$, $p=0.644$, $p=0.996$). There were no differences in these scores within stations (Table 1, Fig. 2).

Table 1 Student notation by the evaluators. Multiple t-tests. *M = men, W = women*

Station number	Evaluated area	Notation	*p*-value
1	Total score (/20)	M: 8.1 ± 0.95; F: 7.52 ± 3.09	0,178
	Medical knowledge (/5)	M: 1.58 ± 1; F: 1.26 ± 0.93	0.054
	Interpersonal skills (/5)	M: 1.98 ± 0.87; F: 1.81 ± 0.9	0.178
	Overall performance (/5)	M: 1.68 ± 0.95; F: 1.43 ± 0.94	0.093
2	Total score (/20)	M: 11.81 ± 2.94; W: 12.27 ± 2.83	0.231
	Medical knowledge (/5)	M: 2.09 ± 0.73; W: 2.29 ± 0.87	0.106
	Interpersonal skills (/5)	M: 2.2 ± 0.78; W: 2.39 ± 0.83	0.106
	Overall performance (/5)	M: 2.14 ± 0.75; W: 2.34 ± 0.88	0.106
3	Total score (/20)	M: 10.85 ± 2.94; W: 10.72 ± 2.81	0.917
	Medical knowledge (/5)	M: 2.06 ± 0.8; W: 2.07 ± 0.78	0.917
	Interpersonal skills (/5)	M: 2.35 ± 0.84; W: 2.41 ± 0.68	0.917
	Overall performance (/5)	M: 2.09 ± 0.8; W: 2.11 ± 0.8	0.917
Average of the 3 stations	Total score (/20)	M: 10.25 ± 3.45; W: 10.23 ± 3.44	0.817
	Medical knowledge (/5)	M: 1.91 ± 0.88; W: 1.87 ± 0.97	0.734
	Interpersonal skills (/5)	M: 2.17 ± 0.84; W: 2.2 ± 0.85	0.644
	Overall performance (/5)	M: 1.97 ± 0.86; W: 1.96 ± 0.95	0.996

M men, *W* women

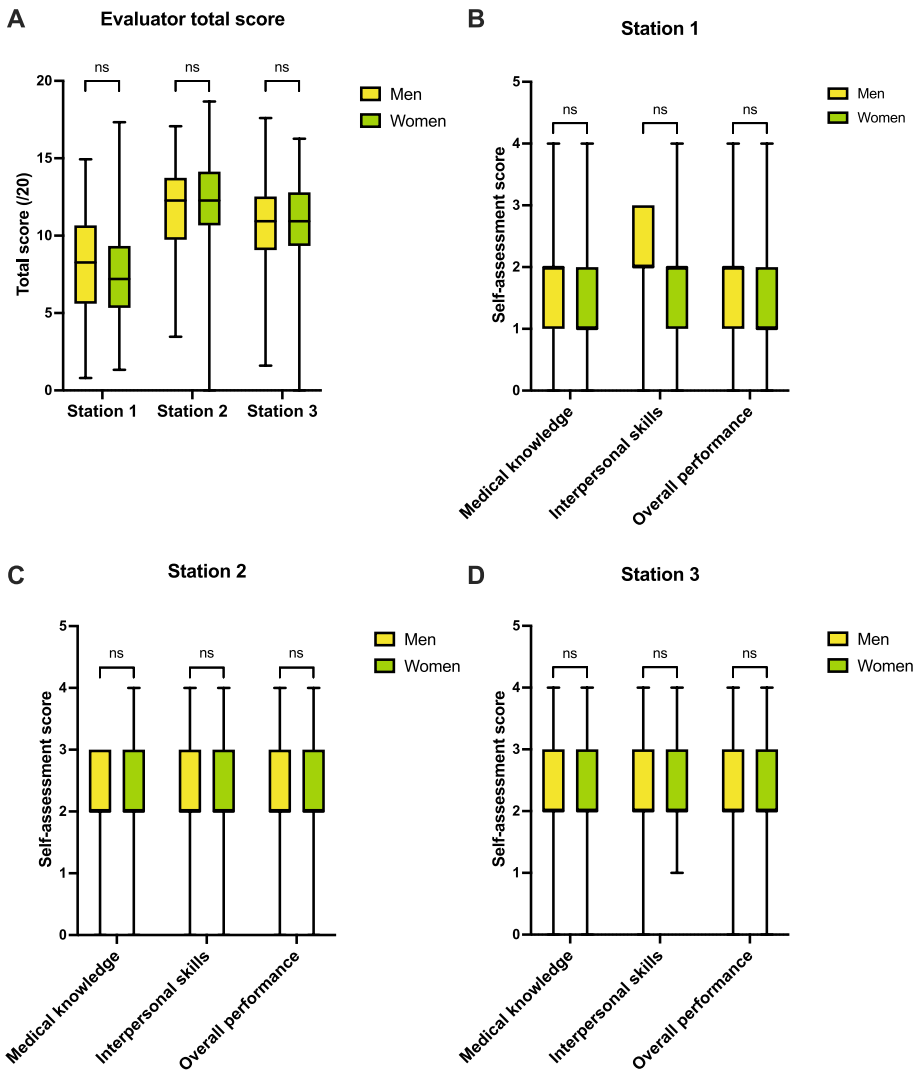

Fig. 2 Student notation by the evaluators depending on gender. Grade out of 20 for each station attributed by the evaluator in each station (**A**). 5-point Likert scales evaluation for stations 1 (**B**), 2 (**C**), and 3 (**D**) on medical knowledge, interpersonal skills, and overall performance. Men are depicted in Yellow, and women in green. Multiple t-tests

No gender difference in students' self-assessments

Self-assessment was completed by 242 students for station 1, 250 for station 2 and 251 for station 3.

Regarding self-assessment, the predicted total score did not differ between women and men: 11.22 ± 3.02 vs. 11.00 ± 3.03 out of 20 ($p = 0.466$). Neither self-attributed scores in any of the 3 Likert scales nor the total score in any stations differed significantly between genders (Table 2). Self-attributed grades also distributed similarly (Figure S4).

The delta between self-assessed-predicted total score and total score by evaluators did not differ between men and women: 0.98 ± 3.41 vs. 0.68 ± 3.42 ($p = 0.296$). Similarly, this delta was similar for stations 1, 2, and 3 ($p = 0.793$, $p = 0.177$, $p = 0.309$, respectively). No difference was observed in either of the 3 Likert scales, with deltas for medical knowledge, interpersonal skills, and overall performance of: 0.73 ± 1.00 vs. 0.64 ± 1.02; 1.02 ± 1.06 vs. 0.93 ± 1.09, and 0.78 ± 0.99 vs 0.69 ± 1.01 for men vs. women respectively, ($p = 0.952$, $p = 0.238$, $p = 0.417$ respectively). Similar results were observed for each station (Table 2, Fig. 3).

Students, regardless of their gender, tend to accurately self-assess

With all deltas very close to zero, both men and women tended to evaluate themselves accurately. However, there was constantly a slight overestimation trend (all deltas

Table 2 Student self-assessment. Multiple t-tests. *M = men, W = women*

Station number	Evaluated area	Self-assessment	*p*-value	Delta	*p*-value
1	Total score (/20)	M: 9.46 ± 3.32, W: 9.22 ± 3.21	0.722	M: 1.47 ± 3.79, W: 1.61 ± 3.68	0.952
	Medical knowledge (/5)	M: 2.16 ± 0.95, W: 2.02 ± 0.8	0.666	M: 0.6 ± 1.1, W: 0.72 ± 1.08	0.816
	Interpersonal skills (/5)	M: 2.9 ± 0.95, W: 2.77 ± 0.90	0.666	M: 0.92 ± 1.13, W: 0.93 ± 1.19	0.952
	Overall performance (/5)	M: 2.3 ± 0.84, W: 2.26 ± 0.81	0.722	M: 0.65 ± 1.01, W: 0.81 ± 1.09	0.816
2	Total score (/20)	M: 11.94 ± 2.61, W: 11.87 ± 2.63	0.864	M: 0.13 ± 3.49, W: -0.49 ± 3.25	0.238
	Medical knowledge (/5)	M: 2.89 ± 0.89, W: 2.81 ± 0.74	0.864	M: 0.8 ± 0.98, W: 0.49 ± 1.05	0.121
	Interpersonal skills (/5)	M: 3.28 ± 0.74, W: 3.32 ± 0.83	0.864	M: 1.04 ± 0.99, W: 0.89 ± 1.13	0.320
	Overall performance (/5)	M: 2.94 ± 0.77, W: 2.91 ± 0.68	0.864	M: 0.8 ± 0.98, W: 0.54 ± 1.02	0.132
3	Total score (/20)	M: 12.23 ± 2.27, W: 11.84 ± 2.43	0.444	M: 1.35 ± 2.75, W: 0.95 ± 2.98	0.417
	Medical knowledge (/5)	M: 2.86 ± 0.68), W: 2.79 ± 0.7	0.444	M: 0.78 ± 0.91, W: 0.71 ± 0.89	0.586
	Interpersonal skills (/5)	M: 3.45 ± 0.68, W: 3.37 ± 0.75	0.444	M: 1.1 ± 1.06, W: 0.96 ± 0.94	0.417
	Overall performance (/5)	M: 2.99 ± 0.63, W: 2.84 ± 0.63	0.397	M: 0.9 ± 0.98, W: 0.73 ± 0.90	0.417
Stations pooled	Total score (/20)	M: 11.22 ± 3.02, W: 11 ± 3.03	0.466	M: 0.98 ± 3.41, W: 0.68 ± 3.42	0.296
	Medical knowledge (/5)	M: 2.64 ± 0.91, W: 2.54 ± 0.83	0.466	M: 0.73 ± 1, W: 0.64 ± 1.02	0.296
	Interpersonal skills (/5)	M: 3.21 ± 0.83, W: 3.16 ± 0.87	0.470	M: 1.02 ± 1.06, W: 0.93 ± 1.09	0.296
	Overall performance (/5)	M: 2.74 ± 0.81, W: 2.67 ± 0.77	0.466	M: 0.78 ± 0.99, W: 0.69 ± 1.01	0.296

M = men, W = women

positive, except total score of women in station 2; Table 2, Fig. 3).

In order to confirm this feature, participants were then categorized into three groups: underestimation, accurate self-assessment, and overestimation. The results indicate that most students accurately assessed themselves in most areas: 78.5% for medical knowledge, 70.5% for interpersonal skills, and 78.1% for overall performance, regardless of gender (Fig. 4). Among those with inaccurate self-assessment, a minority underestimated themselves (for overall performance: 1,2% for men vs. 1% for women, $p = 0.999$), with significantly more students overestimating their capacity independently of gender: 19.6% ($p = 0.0003$) for medical knowledge (21.6% for men vs. 18.8% for women, $p = 0.287$), 28.2% ($p < 0.0001$) for interpersonal skills (30.9% for men vs., 27.1% for women, $p = 0.674$), and 20.9% ($p < 0.0001$) for overall performance (21.9% for men vs. 20.5% for women, $p = 0.956$) (Fig. 4).

Notably, students were less accurate in self-assessing their interpersonal skills and more frequently overestimated their abilities across all stations in this domain.

Overestimation is rarely recurrent

To determine whether some students have a persistent false self-image (i.e., consistently overestimating or underestimating themselves) or if their misjudgments were occasional, we conducted a subanalysis among the 157 women and 74 men who self-evaluated across the three stations. For the same participant, there was no repetition in misjudgment of medical knowledge or overall performance across stations, suggesting that misjudgment is not an inherent trait of a specific individual (relative risk (RR) 1.02 (0.95 – 1.09) $p = 0.661$; Table 3, Fig. 5). However, the risk for a specific student to redundantly overestimate himself concerning the interpersonal skills was higher (RR 1,2 (1.10 – 1.31) $p < 0.0001$; Table 3, Fig. 5).

Conversely, participants consistently misjudged themselves for a given station across the three independent scales (Fig. 5, S4). Collectively, these findings support the hypothesis that students' misjudgments are situational rather than indicative of a persistent tendency to overestimate. Additionally, no gender differences were observed (Figures S5, S6 & S7).

Discussion

Main results

This study observed no significant differences in self-assessment accuracy between male and female medical students during online OSCEs. Both genders demonstrated predominantly accurate self-assessments, with a slight tendency toward overestimation rather than underestimation. Students misjudged their performance inconsistently across the three stations, suggesting situational rather than inherent tendencies. However, interpersonal skills were more frequently overestimated, highlighting an area for improvement in self-perception training.

Comparison to the existing literature

Previous research consistently indicated that women tend to underestimate their skills, while men overestimate or accurately assess theirs [8, 11, 12, 22]. This

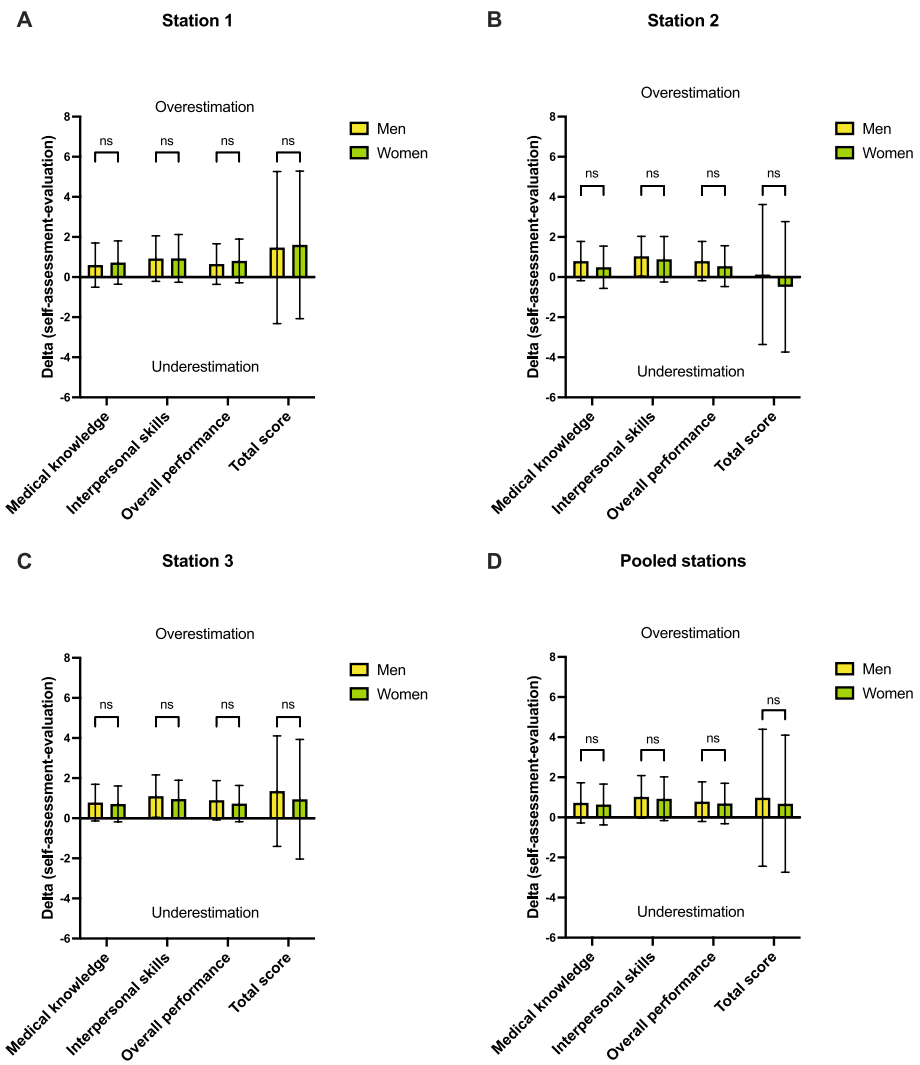

Fig. 3 Difference between self-assessment and evaluation by physicians, depending on gender. For each student, the delta was defined as the difference between the score the student self-assigned and the score from evaluators. Delta on medical knowledge, interpersonal skills, overall performance, and grade out of 20 for stations 1 (**A**), 2 (**B**), and 3 (**C**) and for the 3 stations pooled together (**D**). Men are depicted in Yellow, and women in green. Multiple t-tests

disparity in self-perception appears to intensify over time in medical training. For example, Richmand et al. found no gender differences in self-confidence among first-year students, whereas by the third year, women were significantly more concerned than men about their future abilities and competencies [18]. This trend underscores the growing influence of social and psychological factors on self-assessment performance during medical education [27].

The absence of gender differences in our study might reflect recent societal shifts toward less gendered norms and self-perception. French female medical students may also have benefited from secondary education programs promoting gender equality, which could have fostered increased confidence and engagement in medical careers [28]. Additionally, the majority representation of women among medical students (66% in France in 2020–2021 [29]) and physicians (61% of < 40-year-old physicians and 50,2% of all physicians in France in 2020 [30]) may play a role by boosting self-confidence among female students, potentially reducing disparities in skill assessment.

Gender differences in self-assessment are consistently observed across various exam types, subject areas, and contexts. For instance, in practical, high-pressure scenarios such as intensive care exercises, female first responders were less likely than their male counterparts to assume leadership roles or make critical decisions, suggesting lower confidence in such settings [31, 32].

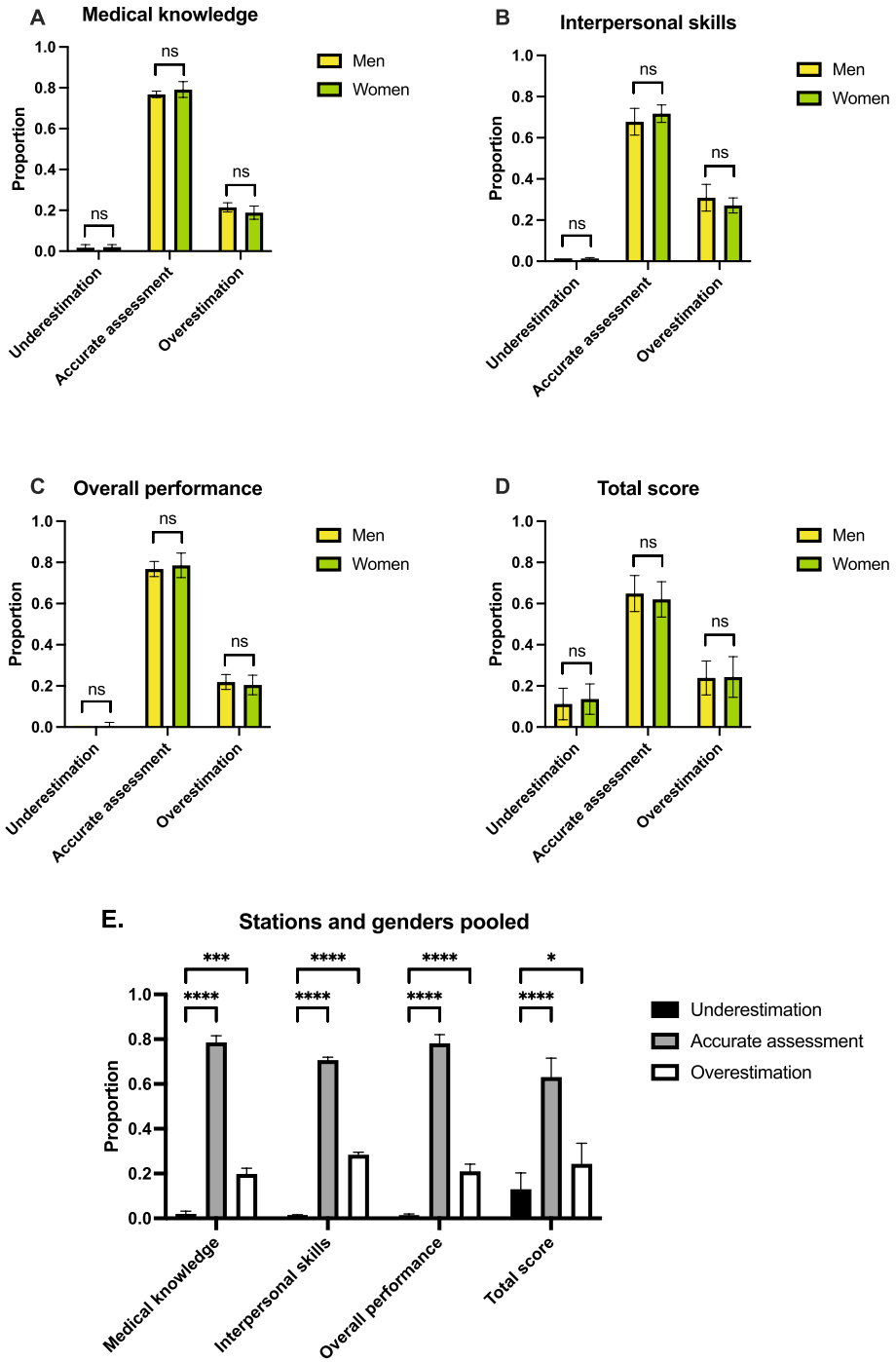

Fig. 4 Proportion of underestimation, accurate assessment and overestimation depending on the skill evaluated. For each station and each Likert scale (medical knowledge (**A**), interpersonal skills (**B**), and overall performance (**C**)), students were divided into 3 groups: underestimation (if the delta was equal or was smaller than -2), accuracy assessment (if the delta was 0 or 1) and overestimation (if delta equal or was greater than 2). The same was done for the total score with accurate assessment defined as a delta equal to or smaller than 3 (**D**). Stations were pooled together. Men are depicted in Yellow, and women in green. E, comparison of underestimation (black) with accurate estimation (grey) or overestimation (white), genders and stations pooled together. 2way ANOVA. * $p < 0.05$, ** $p < 0.01$

Table 3 Comparison of redundant overestimation in medical knowledge, interpersonal skills, and overall performance. Redundant overestimation is defined as at least 2 stations in which a particular student overestimates himself. No redundant overestimation is defined as 0 or 1 station with overestimation. Fisher's exact test

	No redundant overestimation	Redundant overestimation	Relative risk	95% Confidence Interval	*p*
Medical knowledge	206 (89.2%)	25 (10.8%)	-	-	-
Interpersonal skills	172 (74.5%)	59 (25.5%)	1.2	1.100—1.314	< 0.0001
Overall performance	203 (87.9%)	28 (12.1%)	1.02	0.9488—1.086	0.661

Redundant overestimation is defined as at least 2 stations in which a particular student overestimates himself. No redundant overestimation is defined as 0 or 1 station with overestimation. Fisher's exact test

Similarly, in tasks requiring manual skills—such as central venous catheter insertion [33], pre-clinical operative dentistry exams, [34, 35] or surgical procedures [36]—overestimation was more prevalent among male students, while female students demonstrated more accurate or conservative self-assessments and lower confidence level. Consistently, during standardized patient interaction exams, female medical students appeared significantly less confident than their male peers, as perceived by independent observers, despite demonstrating equivalent or superior performance [8]. To the best of our knowledge, this study is the first to examine gender differences in the context of an online OSCE. The lack of significant gender differences in self-assessment accuracy could be linked to the unique nature of this setting, which emphasized semiology rather than practical or manual skills, potentially reducing some of the gender disparities commonly observed in hands-on, in-person assessments.

Limitations

This study has limitations. Some students chose not to participate in this optional OSCE session, and among those who did, some declined to take part in this study. These factors may lead to an overrepresentation of students who felt more confident about the test. However, the data collection was robust, with only 11 students excluded due to technical issues or non-compliance. Moreover, participants were unaware of the study's objective before their involvement, and none withdrew consent after being informed, minimizing potential bias.

The evaluators' gender influence, which remains a debated issue [23, 37], was not assessed as evaluators worked in pairs. This setup allows us to minimize the impact of such a bias and to assume that there is no difference in evaluation between the evaluators.

The generalizability of this study is also limited. Conducted at a single university in an urban, the findings may not apply to students from rural or socioeconomically disadvantaged backgrounds. Paris Cité University's predominantly privileged student demographic could skew results, as role models and access to medical professionals within their social circles may enhance confidence [38]. Furthermore, the online OSCEs format, with students participating from their homes, may have reduced their stress levels and improved self-assessment accuracy compared to in-person assessments. Indeed, online OSCEs introduced during the COVID-19 pandemic have been shown to reduce anxiety levels as students participate from familiar environments, which could reduce the perceived seriousness of the evaluation [39]. However, other research comparing online and live OSCEs has demonstrated that both formats are generally well-received, with minimal differences in perceived exam quality, though live OSCEs often feel more authentic and allow for better time management [40]. Other findings indicate that virtual formats are comparable in terms of skill mastery evaluation but may limit the assessment of manual skills [41]. Finally, even though most participants in virtual OSCE sessions reported improved confidence in their skills in history-taking, communication, and data interpretation, Grover et al. showed that among participants and examiners who had also participated in in-person OSCE examinations, the majority reported finding virtual OSCE sessions as engaging and interactive as in-person teaching [42].

Another important point is that our study was conducted in a training setting, where the results had no impact on the students' academic progression. Consequently, the level of stress generated by this type of OSCE was likely lower than that of a high-stakes exam required for advancing in their studies.

Additionally, the involvement of 130 evaluators may have introduced variability in scoring due to differences in professional background, experience, or subjective judgment. While all evaluators participated in preparatory sessions and worked in pairs to mitigate bias, variability remains an inherent challenge in large-scale assessments.

Finally, only one student in the cohort identified as non-binary (0.4%), potentially underrepresenting this

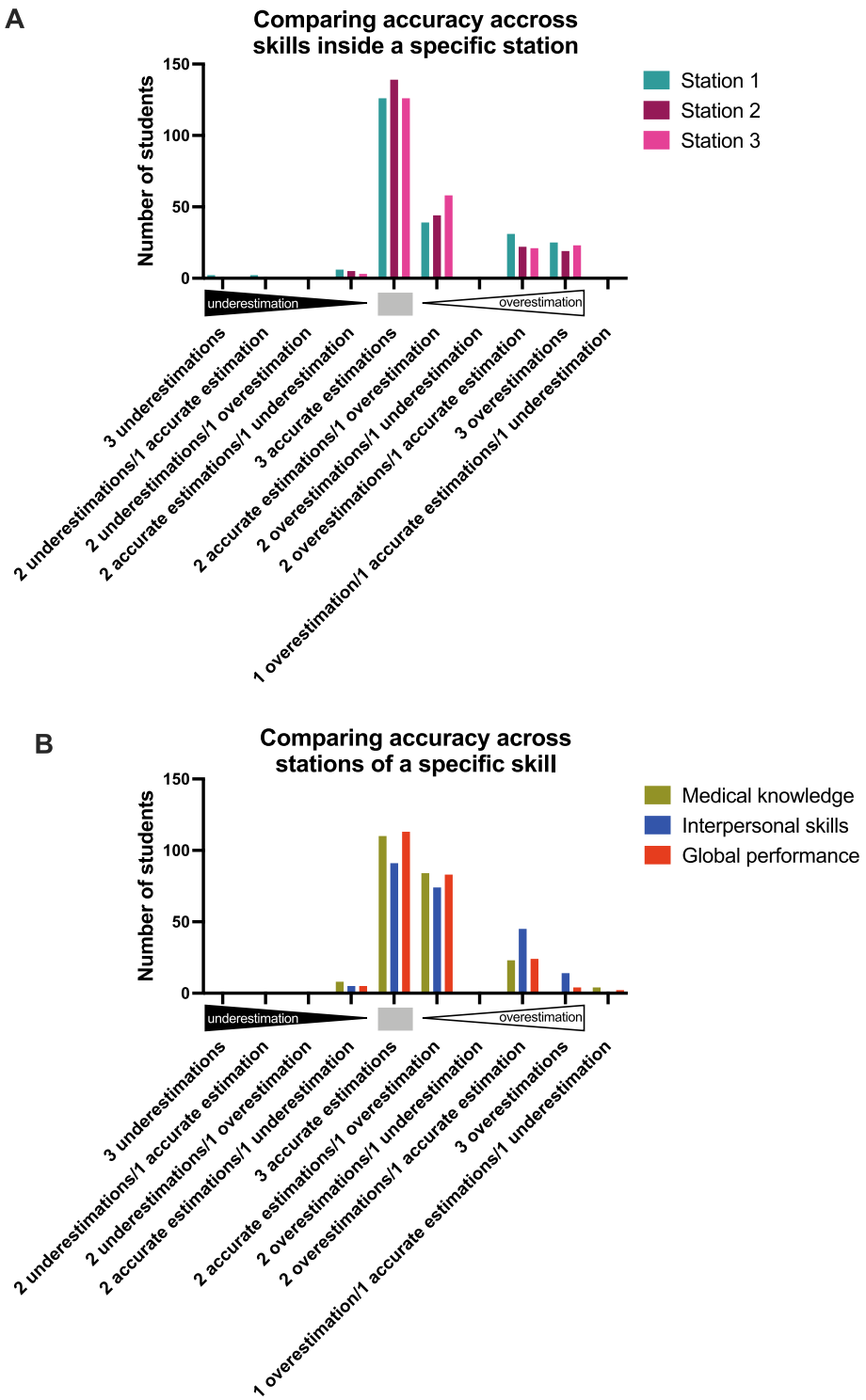

Fig. 5 Consistency of misjudgment across stations and skills. This analysis included 231 students (157 women and 64 men) who completed all skills across the three stations. Overestimation was defined as a delta ≥ 2, accurate assessment as a delta of 0 or 1, and underestimation as a delta ≤ -2. Consistency of misjudgment was analyzed across stations for each skill (**A**) and across skills for each station (**B**)

group. As Velin et al. suggest, this proportion may be closer to 1%, but data on non-binary students in medical education remains limited [43].

Further research

Future studies should confirm these results in more diverse populations, including multiple universities and regions with varying socioeconomic conditions. Factors such as parental background, early education, and access to role models should be explored to better understand disparities in self-assessment. Further investigation is also needed to examine the impact of in-person versus remote OSCE formats on self-assessment accuracy. Particular attention should be given to the recurring overestimation of interpersonal skills. Incorporating non-healthcare professionals or simulated patients as evaluators might provide a more accurate assessment of relational competencies. Additional studies are also needed to gather data on underrepresented groups, including non-binary students, to gain insights into their self-assessment dynamics.

Finally, future research could reduce evaluator-related variability by employing smaller, highly trained groups of evaluators and conducting inter-rater reliability analyses to ensure scoring consistency.

Conclusion

This study suggests a potential shift towards gender-equitable self-perceptions among medical students, contrasting with previous findings of persistent disparities. If this conclusion carries hope, further more in-depth research is needed with more diversity of the study population, and more socio-economics parameters investigated to fully understand this dynamic.

Abbreviation

OSCEs Objective Structured Clinical Examinations

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-024-06573-x.

Supplementary Material 1: Figure S1. Ventricular tachycardia station: Electrocardiogram Interpretation and Immediate Management of Chest Discomfort (A. Clinical vignette of the student, B. Assessor's evaluation grid, C. Instructions for the standardized healthcare professional (SHP)). Figure S2. Diarrhea station: Evaluation and Management of Acute Diarrhea (A. Clinical vignette of the student, B. Assessor's evaluation grid, C. Instructions for the standardized patient (SP)). Figure S3. Asthma station: Assessment and Diagnostic Strategy for Chronic Cough (A. Clinical vignette of the student, B. Assessor's evaluation grid, C. Instructions for the standardized patient (SP)). Figure S4. Consistency of Misjudgment Across Stations and Scales. Each line represents a student (women or men) for whom we have data for all 3 stations and all skills. Each column is a station (A, B, and C) or a skill (D, E, and F). In each situation, overestimation is depicted in white (if the delta was equal or was greater than 2), accurate assessment in gray (if the delta was 0 or 1), and underestimation in black (if the delta was equal or was smaller than -2). Figure S5. Distribution of grades predicted by the students. Men are depicted in Yellow, and women in green. Figure S6. Consistency of Misjudgment Across Stations Depending on Gender. Each line represents a student, women (A, B, and C) or men (D, E, and F) for whom we have data for all 3 stations and all skills. Each column is a skill. In each situation, overestimation is depicted in white (if the delta was equal or was greater than 2), accurate assessment in gray (if the delta was 0 or 1), and underestimation in black (if the delta was equal or was smaller than -2). Figure S7. Consistency of Misjudgment Across Scales Depending on Gender. Each line represents a student, women (A, B, and C) or men (D, E, and F) for whom we have data for all 3 stations and all skills. Each column is a station. In each situation, overestimation is depicted in white (if the delta was equal or was greater than 2), accurate assessment in gray (if the delta was 0 or 1), and underestimation in black (if the delta was equal or was smaller than -2).

Supplementary Material 2: Table S1. Likert Scales for evaluation of Medical knowledge and Overall Performance. Table S2. Likert Scales for Evaluation of Interpersonal Skills. Table S3: Grid for ventricular tachycardia station evaluation. Table S4: Grid for diarrhea station evaluation. Table S5: Grid for asthma station evaluation. Table S6: Complete student survey. Table S7. Evaluator functions.

Supplementary Material 3: Table S8. Raw dataset.

Acknowledgements

The authors thank all the students and all the physicians of Paris Cité University who participated in the study for their valuable contributions. LIL thanks Melissa Ait Said for valuable discussions.

Clinical trial number

Not applicable.

Authors' contributions

LIL, AF, DB, VMF, CC, JK, YN, CDT, NB and TM designed and conceptualized the study. LIL collected all data. LIL and SB analyzed data and wrote the initial manuscript. All authors reviewed the manuscript. LIL, AF supervised and coordinate the study.

Funding

None.

Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information files (Table S8).

Declarations

Ethics approval and consent to participate

That work was carried out in accordance with the Declaration of Helsinki, including, but not limited to the anonymity of participants being guaranteed and the informed consent of participants being obtained. The Institutional Review Board of Paris Cité University approved this study (IRB No: 00012023–10).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Radiology, Hôpital Necker, AP-HP, Université Paris Cité, Paris F-75015, France. [2]Sorbonne University, CNRS UMR 7371, INSERM U 1146, Laboratoire d'Imagerie Biomédicale, Paris F-75006, France. [3]Center for Transplantation Sciences, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [4]Emergency Department, Hôpital Bichat-Claude Bernard, AP-HP, Université Paris Cité, Paris F-75018, France. [5]Virology Department, Université Paris Cité, Hôpital Bichat-Claude Bernard, AP-HP, IAME, INSERM UMR_1137, Paris F-75018, France. [6]Department of Nutrition, Hôpital Européen Georges Pompidou, AP-HP, Université Paris Cité, Paris F-75015,

France. [7]Université Paris Cité, ECEVE, UMR 1123, Inserm, Paris F-75010, France. [8]Department of Internal Medicine, Hôpital Beaujon, AP-HP, Université Paris Cité, Clichy F-92110, France. [9]Center for Epidemiology and Statistics, CRESS, INSERM U1153, Université Paris Cité, Paris F-75006, France. [10]Medical and Infectious Diseases ICU, Hôpital Bichat Claude Bernard, AP-HP, Université Paris Cité, Paris F-75018, France. [11]Decision Sciences in Infectious Diseases Control and Care, UMR 1137-IAME-INSERM Université Paris Cité, Paris F-75018, France. [12]Department of Digestive, Hepatobiliary and Endocrine Surgery, Hôpital Cochin, AP-HP, Université Paris Cité, Paris F-75014, France. [13]Department of Dermatology, INSERM U1016 Institut Cochin, Hôpital Cochin AP-HP, Université Paris Cité, Paris F-75014, France. [14]Departement of Vascular Medicine, INSERM U970 PARCC, Hôpital Européen Georges-Pompidou, AP-HP, Université Paris Cité, Paris F-75015, France. [15]Université Paris Cité, INSERM UMR1149, CRI, Paris F-75018, France. [16]Anesthesia and Intensive Care Department, Hôpital Louis Mourier, APHP, Colombes F-92700, France. [17]Université Paris Cité, INSERM, PARCC, Paris F-75015, France. [18]Department of Tumor and Cancer Genomic Medicine, Federation of Genetics and Genomic Medicine, Hôpital Européen Georges Pompidou, APHP, Paris F-75015, France. [19]Department of General Pediatrics, Infectious Diseases and Internal Medicine, Hôpital Robert Debré, AP-HP, Université Paris Cité, Paris F-75019, France. [20]Department of Infectious and Tropical Diseases, Hôpital Saint-Louis Et Lariboisière, AP-HP, Université Paris Cité, 1 Avenue Claude Vellefaux, Paris F-75010, France.

Received: 10 September 2024 Accepted: 19 December 2024
Published online: 30 December 2024

References

1. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. BMJ. 1975;1(5955):447–51.
2. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ. 1979;13(1):41–54.
3. Patrício MF, Julião M, Fareleira F, Carneiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? Med Teach. 2013;35(6):503–14.
4. Fischer MA, Kennedy KM, Durning S, Schijven MP, Ker J, O'Connor P, et al. Situational awareness within objective structured clinical examination stations in undergraduate medical training - a literature search. BMC Med Educ. 2017;17(1):262.
5. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak. 2016;16(1):138.
6. Hainguerlot M, Gajdos T, Vergnaud JC, De Gardelle V. How overconfidence bias influences suboptimality in perceptual decision making. J Exp Psychol Hum Percept Perform. 2023;49(4):537–48.
7. Blanch-Hartigan D. Medical students' self-assessment of performance: Results from three meta-analyses. Patient Educ Couns. 2011;84(1):3–9.
8. Blanch DC, Hall JA, Roter DL, Frankel RM. Medical student gender and issues of confidence. Patient Educ Couns. 2008;72(3):374–81.
9. Wilson DA, Warton C, Louw GJ. Stress, Anxiety And Academic Performance Among Medical Students At The University Of Cape Town. South Afr J Child Adolesc Ment Health. 1998;10(1):23–31.
10. Graf J, Smolka R, Simoes E, Zipfel S, Junne F, Holderried F, et al. Communication skills of medical students during the OSCE: Gender-specific differences in a longitudinal trend study. BMC Med Educ. 2017;17(1):75.
11. Colbert-Getz JM, Fleishman C, Jung J, Shilkofski N. How Do Gender and Anxiety Affect Students' Self-Assessment and Actual Performance on a High-Stakes Clinical Skills Examination?: Acad Med. 2013;88(1):44–8.
12. Madrazo L, Lee CB, McConnell M, Khamisa K. Self-assessment differences between genders in a low-stakes objective structured clinical examination (OSCE). BMC Res Notes. 2018;11(1):393.
13. Knorr M, Meyer H, Sehner S, Hampe W, Zimmermann S. Exploring sociodemographic subgroup differences in multiple mini-interview (MMI) performance based on MMI station type and the implications for the predictive fairness of the Hamburg MMI. BMC Med Educ. 2019;19(1):243.
14. Afshar K, Wiese B, Stiel S, Schneider N, Engel B. Perceived stress and study-related behavior and experience patterns of medical students: a cross-sectional study. BMC Med Educ. 2022;22(1):122.
15. Carney DR, Colvin CR, Hall JA. A thin slice perspective on the accuracy of first impressions. J Res Personal. 2007;41(5):1054–72.
16. Egloff B, Schmukle SC. Gender differences in implicit and explicit anxiety measures. Personal Individ Differ. 2004;36(8):1807–15.
17. Lang NP, Rowland-Morin PA, Coe NPW. Identification of communication apprehension in medical students starting a surgery rotation. Am J Surg. 1998;176(1):41–5.
18. Richman JA, Flaherty JA. Gender differences in medical student distress: contributions of prior socialization and current role-related stress. Soc Sci Med 1982. 1990;30(7):777–87.
19. Minter RM, Gruppen LD, Napolitano KS, Gauger PG. Gender differences in the self-assessment of surgical residents. Am J Surg. 2005;189(6):647–50.
20. Lind DS, Rekkas S, Bui V, Lam T, Beierle E, Copeland EM. Competency-Based Student Self-Assessment on a Surgery Rotation. J Surg Res. 2002;105(1):31–4.
21. Yudkowsky R, Downing SM, Ommert D. Prior experiences associated with residents' scores on a communication and interpersonal skill OSCE. Patient Educ Couns. 2006;62(3):368–73.
22. Blanch-Hartigan D. Medical students' self-assessment of performance: Results from three meta-analyses. Patient Educ Couns. 2011;84(1):3–9.
23. Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J. Science faculty's subtle gender biases favor male students. Proc Natl Acad Sci. 2012;109(41):16474–9.
24. Toews JA, Lockyer JM, Dobson DJ, Simpson E, Brownell AK, Brenneis F, et al. Analysis of stress levels among medical students, residents, and graduate students at four Canadian schools of medicine: Acad Med. 1997;72(11):997–1002.
25. Bouzid D, Mirault T, Ghazali A, Muller L, Casalino E, Peiffer Smadja N, et al. Feasibility of large-scale eOSCES: the simultaneous evaluation of 500 medical students during a mock examination. Med Educ Online. 2022;27(1):2084261.
26. Graves L, Lalla L, Young M. Evaluation of perceived and actual competency in a family medicine objective structured clinical examination. Can Fam Physician Med Fam Can. 2017;63(4):e238–43.
27. Carlin BA, Gelb BD, Belinne JK, Ramchand L. Bridging the gender gap in confidence. Bus Horiz. 2018;61(5):765–74.
28. Ministère de l'Education Nationale et de la Jeunesse. Égalité entre les filles et les garçons. Available from: https://www.education.gouv.fr/egalite-entre-les-filles-et-les-garcons-9047. [Cited 2024 Aug 23].
29. Insee Références. Femmes et hommes, l'égalité en question. 2022. Available from: https://www.insee.fr/fr/statistiques/6047727?sommaire=6047805.
30. Bouet P, Gerard-Varet JF. ATLAS DE LA DÉMOGRAPHIE MÉDICALE EN FRANCE, SITUATION AU 1er JANVIER 2020. Ordre National de Médecins; Available from: https://www.conseil-national.medecin.fr/sites/default/files/external-package/analyse_etude/1grhel2/cnom_atlas_demographie_medicale_2020_tome1.pdf.
31. Hochstrasser SR, Amacher SA, Tschan F, Semmer NK, Becker C, Metzger K, et al. Gender-focused training improves leadership of female medical students: a randomised trial. Med Educ. 2022;56(3):321–30.
32. Olson EM, Sanborn DM, Dyster TG, Kelm DJ, Murray SG, Santhosh L, et al. Gender Disparities in Critical Care Procedure Training of Internal Medicine Residents. Sch. 2023;4(2):164–76.
33. Solberg M, Wong A, Ikejiani S, et al. Comparing Male and Female Resident Physicians in Central Venous Catheter Insertion Self-confidence and Competency: A Retrospective Cohort Study. J Gen Intern Med. 2024. https://doi.org/10.1007/s11606-024-08982-6.
34. Sedky RAF, Ben Dor B, Mustafa DS, Galal MM, Nour KA, Anwar MN, et al. Self-assessment skills of undergraduate students in operative dentistry: preclinical performance and gender. Dent Med Probl. 2024. 27;(6):0–0.
35. Javed MQ, Bhatti UA. Students' performance in clinics and self-perceived Confidence in performing Endodontic procedures: a correlation study. Pak J Med Sci. 2023;39(1):203–8.
36. Stanek K, Phillips N, Staffa SJ, Saldanha FYL, Rogers-Vizena CR. Gender differences in plastic surgery trainee confidence: a pilot analysis during cleft lip simulation. Plast Reconstr Surg Glob Open. 2023;11(12):e5428.
37. 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.

38. Khan R, Apramian T, Kang JH, Gustafson J, Sibbald S. Demographic and socioeconomic characteristics of Canadian medical students: a cross-sectional study. BMC Med Educ. 2020;20(1):151.
39. Tuononen T, Karaharju-Suvanto T, Lahti S, Hytönen H, Näpänkangas R. Dental students' perceptions of simultaneous live and online OSCEs during the COVID-19 pandemic. Eur J Dent Educ Off J Assoc Dent Educ Eur. 2024;28(2):408–15.
40. Loda T, Erschens RS, Nevins AB, Zipfel S, Herrmann-Werner A. Perspectives, benefits and challenges of a live OSCE during the COVID-19 pandemic in a cross-sectional study. BMJ Open. 2022;12(6):e058845.
41. Desai PV, Howell HB, McGrath M, Ramsey R, Lebowitz J, Trogen B, et al. Zoom objective structured clinical exams: virtually the same as the real thing? Acad Pediatr. 2023;23(2):483–8.
42. Grover S, Pandya M, Ranasinghe C, Ramji SP, Bola H, Raj S. Assessing the utility of virtual OSCE sessions as an educational tool: a national pilot study. BMC Med Educ. 2022;22(1):178.
43. Velin L, Chew MS, Pompermaier L. Discrimination in an "equal country"—a survey amongst Swedish final-year medical students. BMC Med Educ. 2022;22(1):503.

Publisher's Note

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