

Resident–faculty differences in global performance scores in otolaryngology workplace-based assessment: a national retrospective observational study

Okki Dhona Laksmita

National Taiwan University

Yi-Jia Chen

Cardinal Tien Hospital and Fu Jen Catholic University

Ching-Lin Hsieh

National Taiwan University

Ya-Hui Wang

Cardinal Tien Hospital and Fu Jen Catholic University

Chun-Hou Liao

Cardinal Tien Hospital and School of Medicine, Fu Jen Catholic University

Mingchih Chen

Fu Jen Catholic University

Jeng-Wen Chen

086365@mail.fju.edu.tw

Cardinal Tien Hospital and Fu Jen Catholic University

Research Article

Keywords: workplace-based assessment (WBA), entrustable professional activities (EPAs), resident self-assessment, faculty assessment, score discrepancy, otolaryngology residency

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10932533/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Resident self-assessment is integral to workplace-based assessment (WBA), but resident–faculty score differences remain unclear in national entrustable professional activity (EPA)-based systems. We quantified these differences and examined variation by training level and case complexity in Taiwan's national otolaryngology WBA platform.

Methods

We retrospectively analyzed paired resident and faculty ratings from the E-MyWay platform. Discrepancy was defined as resident self-score minus faculty score on a 1–10 scale. Mixed-effects linear and logistic models included random intercepts for residents and faculty raters and adjusted for assessment and contextual factors.

Results

The analysis included 47,346 records from 466 residents, 447 faculty raters, and 12 EPAs. The mean discrepancy was – 0.6 points. Exact agreement occurred in 41.8% of records, and 80.2% were within one point. Resident self-scores were lower than faculty scores in 48.5% of records and higher in 9.7%. Higher training level was associated with lower odds of resident-higher discrepancy (OR 0.81, 95% CI 0.76–0.85) and resident-higher discrepancy of at least two points (OR 0.68, 95% CI 0.61–0.76). Advanced/non-routine cases were associated with lower odds of within-one-point agreement (OR 0.84, 95% CI 0.79–0.90) and higher odds of resident-lower discrepancy of at least two points (OR 1.18, 95% CI 1.10–1.27).

Conclusions

Residents generally assigned slightly lower scores than faculty raters, although most paired ratings differed by no more than one point. Score differences varied by training level and case complexity and may help identify encounters warranting discussion and feedback.

Background

Competency-based medical education (CBME) increasingly relies on workplace-based assessment (WBA) to generate repeated observations of trainee performance in authentic clinical settings. These assessments are expected not only to document performance, but also to guide feedback, reflection, and developmental decision-making [1, 2]. Entrustable professional activities (EPAs) operationalize this approach by linking competencies to observable clinical work and supervision decisions [3]. As EPA-based WBA systems expand, paired resident and faculty ratings are becoming a readily available source of information about how trainees and supervisors judge the same performance encounter.

The meaning of these paired ratings, however, is not straightforward. Resident self-assessment is important for self-regulated learning, engagement with feedback, and lifelong professional development. Earlier research often evaluated self-assessment as an accuracy judgment against an external rating and found limited correspondence [4–6]. More recent scholarship conceptualizes self-assessment as an informed, developmental process in which learners integrate internal judgments with external feedback, performance standards, and contextual cues [7, 8]. From this perspective, a resident–faculty score difference should not automatically be interpreted as resident error. It may instead reflect different reference standards, perceptions of uncertainty, expectations for the learner's stage, or the way each rater uses the available scale.

Existing evidence is fragmented and has produced inconsistent findings. Studies in surgical and procedural settings have reported lower resident self-ratings than faculty ratings [9, 10], whereas formative WBA studies have found closer alignment or differences of limited practical magnitude [11]. Milestone-based research further suggests that discrepancy patterns vary across learner characteristics, competency domains, and specialties [12]. Most available studies are limited to particular tasks, programs, or settings. It therefore remains unclear whether a consistent resident–faculty pattern is observable across an entire national WBA system and whether that pattern changes with training level, case complexity, or assessment context. This uncertainty limits the appropriate interpretation and educational use of paired ratings.

A national EPA-based platform offers a distinctive opportunity to address this gap. Taiwan's E-MyWay platform supports otolaryngology residency assessment across clinical settings and EPAs [13]. For each encounter, residents record a self-score and reflection, while faculty raters record a global performance score, narrative feedback, and an entrustment-supervision judgment. The resulting data allow resident and faculty perspectives to be compared at scale while accounting for repeated assessments by the same residents and faculty raters. Such a comparison does not establish either rating as the true measure of performance; rather, it characterizes the direction, magnitude, and context of differences between two assessment perspectives.

Clarifying these patterns has practical value for WBA design and feedback. If score differences are generally small and predictable, paired ratings may help educators distinguish routine variation from encounters that warrant further discussion. If differences vary with training stage or clinical complexity, feedback and interpretation may need to be tailored to the demands of the encounter. Training level and case complexity are particularly relevant because experience, self-monitoring, standards, and responses to uncertainty may develop or shift across residency [14–16]. Gender-related factors may also shape self-rating, faculty assessment, and feedback processes, although observational patterns cannot identify their underlying mechanisms [4, 16–18].

Accordingly, the primary aim of this study was to quantify the direction and magnitude of resident–faculty differences in global performance scores within a national otolaryngology WBA platform. Secondary analyses examined whether these differences varied by training level and case complexity. Exploratory analyses examined variation by resident gender, faculty rater gender, and resident–faculty gender pairing after accounting for assessment-task and contextual factors. By examining paired ratings across a national system, this study provides evidence to support more defensible interpretation and educational use of resident–faculty score differences in WBA.

Methods

Ethical statement

The Institutional Review Board of Cardinal Tien Hospital approved this study (CTH-114-2-1-039). All participants provided electronic informed consent before first accessing the E-MyWay platform. Data were de-identified before analysis.

Study design and reporting framework

We conducted a retrospective observational study using routinely collected, encounter-level data from Taiwan's national otolaryngology WBA platform [13, 19–21]. The unit of analysis was an assessment encounter containing paired resident and faculty ratings. The study is reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement [22].

Data source, setting, and eligibility criteria

Data were obtained from E-MyWay, Taiwan's national EPA-based WBA platform for otolaryngology residency training [13, 19–21]. Assessments were grouped into four periods: Period 1, August 2022–July 2023; Period 2, August 2023–July 2024; Period 3, August 2024–July 2025; and Period 4, August 2025–May 2026. For each formative assessment encounter, residents recorded a self-score and reflection, while faculty raters provided a global performance score, narrative feedback, and an entrustment-supervision rating. Faculty raters could view the resident's self-score when completing their assessment; therefore, the paired resident and faculty scores were not collected under blinded or fully independent conditions. This study analyzed paired resident and faculty scores on the 1–10 global rating scale.

Eligible records contained valid resident and faculty identifiers, an R1–R5 training level, an EPA title, and paired scores within the 1–10 range. Records not meeting these criteria were excluded. EPA titles were used to characterize assessment context and define the EPA-group covariate.

Measures and variables

Resident self-scores and faculty global performance scores were paired within each WBA encounter and recorded on the same 1–10 scale. Training level was categorized from R1 to R5. The 12 EPA titles were grouped into four investigator-defined task-context categories: early (EPA01–05), core (EPA07–09), delayed (EPA06, 10, 11), and presentation (EPA12). These categories reflected the developmental structure of Taiwan's otolaryngology EPA framework and were used solely as covariates; they did not constitute an official EPA taxonomy or entrustment outcomes.

Additional variables included clinical setting, case complexity, hospital level, region, assessment period, resident gender, and faculty rater gender. Resident–faculty gender pairing combined the recorded genders of the resident and faculty rater within each encounter and was examined as an exploratory variable to assess whether the continuous resident–faculty discrepancy score differed across gender combinations.

Outcomes

The primary outcome was the resident–faculty discrepancy score, calculated as the resident self-score minus the faculty global performance score within each WBA encounter. Secondary categorical outcomes were derived from this score: overestimation (discrepancy >0), underestimation (discrepancy <0), exact agreement (discrepancy = 0), within-one-point agreement (absolute discrepancy ≤1 point), large overestimation (discrepancy ≥+2 points), and large underestimation (discrepancy ≤−2 points). The two-point threshold identified discrepancies larger than minor adjacent-score differences on the 1–10 scale. In all definitions, faculty scores served as the comparator and were not treated as an error-free measure of true clinical performance.

Covariates

Training level and case complexity were used to examine whether the main resident–faculty discrepancy pattern varied across educational stage and clinical complexity. EPA group, clinical setting, hospital level, region, and assessment period were included as task-context or adjustment covariates to account for differences in assessment conditions. Resident gender and faculty rater gender were included as exploratory gender-related variables in the main adjusted models, whereas resident–faculty gender pairing was examined separately in an exploratory model for the continuous discrepancy-score outcome.

Statistical analysis

Analyses were conducted using R, version 4.5.3 [23]. Mixed-effects models were fitted using the lme4 package [24], and *p* values for the linear mixed-effects model were obtained using the lmerTest package [25].

Descriptive analyses were conducted to summarize resident self-scores, faculty scores, discrepancy scores, and absolute discrepancy scores. Continuous variables were summarized using means and standard deviations. Categorical variables were summarized using frequencies and percentages.

Overall discrepancy patterns were described by calculating the proportion of records in each discrepancy category. These outcomes were summarized overall and by training level, EPA group, case complexity, resident gender, and faculty rater gender.

For multivariable analysis, a mixed-effects linear regression model was used to examine discrepancy score as a continuous outcome. Because WBA records were clustered within residents and faculty raters, the model included random intercepts for resident and faculty rater. Fixed effects included training level, EPA group, clinical setting, case complexity, hospital level, region, assessment period, resident gender, and faculty rater gender. Training level was modeled as an ordinal variable from R1 to R5 to estimate the association per one-level increase in training level.

Secondary outcomes were examined using mixed-effects logistic regression models with the same random-effects and fixed-effects structure. These binary outcomes included overestimation, underestimation, within-one-point agreement, large overestimation, and large underestimation. Exact agreement was summarized descriptively but was not included as a separate adjusted logistic outcome. Model results were reported as regression coefficients with 95% confidence intervals for the continuous discrepancy outcome and as odds ratios with 95% confidence intervals for binary outcomes.

Sensitivity analyses were conducted to examine whether the main adjusted findings were robust to key analytic specifications. First, the adjusted models were repeated with training level modeled as a categorical variable, using R1 as the reference group, rather than as an ordinal R1–R5 variable. Second, the adjusted models were repeated using the 12 individual EPA titles instead of the four investigator-defined EPA groups. Both sensitivity analyses retained the same random-effects structure.

For the continuous discrepancy-score outcome, an exploratory model was also conducted in which resident–faculty gender pairing replaced the separate resident gender and faculty rater gender variables, with male resident–male faculty rater pairs as the reference group.

Statistical tests were two-sided. Because of the large number of WBA records, results were interpreted with emphasis on effect estimates and 95% confidence intervals rather than *p* values alone.

Missing data handling

Missingness was assessed for all analytic variables, and the primary analytic sample was defined by the eligibility criteria described above. Adjusted mixed-effects models used complete-case analysis. Missing covariate data were minimal: two records with missing case-complexity data were excluded from the adjusted models, and all other covariates were complete in the final analytic sample.

Results

Analytic sample and assessment-record characteristics

The source dataset contained 48,339 workplace-based assessment records. After applying the eligibility criteria, the final analytic sample included 47,346 paired WBA records covering 12 EPAs, contributed by 466 residents and 447 faculty raters from 35 otolaryngology residency training programs across Taiwan. Complete data for all variables included in the adjusted mixed-effects models were available for 47,344 assessments; two records with missing case complexity were excluded from these models.

Assessment-record characteristics are summarized in Table 1. Records were distributed across all training levels, with slightly fewer from R5, and across all EPA groups, with most from the early/basic, core disease-management, and delayed/advanced groups. Approximately one-third of encounters involved advanced/non-routine cases, and male residents and male faculty raters accounted for the majority of records.

Overall resident–faculty score discrepancy

Descriptive discrepancy outcomes are summarized in Table 2. The mean discrepancy score was -0.6 points on the 1–10 scale, indicating that resident self-scores were, on average, lower than faculty scores. The 95% limits of agreement ranged from -2.89 to 1.65 (Fig. 1). Underestimation occurred in 48.5% of records and overestimation in 9.7%; exact agreement occurred in 41.8%, and 80.2% of paired scores were within one point. Large overestimation and large underestimation occurred in 2.1% and 17.7% of records, respectively (Additional file 1: Supplementary Table S1).

Discrepancy patterns by training level and clinical context

Both resident self-scores and faculty scores increased with training level, from 7.0 to 7.8 and from 7.5 to 8.5 between R1 and R5, respectively, and faculty scores exceeded resident self-scores at every level (Table 2).

With increasing training level, overestimation decreased from 13.9% (R1) to 7.0% (R5) and underestimation increased from 43.8% to 52.1%, whereas exact agreement remained stable (40.5%–44.0%; Fig. 2).

Compared with basic/routine cases, advanced/non-routine cases showed a more negative mean discrepancy score (-0.8 vs. -0.6), lower within-one-point agreement (75.5% vs. 82.4%), and more frequent large underestimation (22.5% vs. 15.5%; Additional file 1: Supplementary Table S1). Mean discrepancy scores were similar across EPA groups (-0.6 to -0.7).

Exploratory gender-related descriptive patterns

Records involving female residents showed slightly more negative mean discrepancy scores (-0.7 vs. -0.6) and more frequent underestimation (52.4% vs. 47.0%) than records involving male residents (Table 2). A parallel pattern was observed for faculty rater gender: assessments completed by female faculty

raters showed more negative mean discrepancy scores (−0.7 vs. −0.6) and more frequent underestimation (51.6% vs. 47.9%) than those completed by male faculty raters. These patterns were examined further in the adjusted models.

Adjusted mixed-effects models

Adjusted estimates for training level, case complexity, and the exploratory gender-related variables are shown in Table 3; full fixed-effect estimates for all covariates are provided in Additional file 1: Supplementary Table S2.

In the fully adjusted models accounting for clustering by resident and faculty rater, training level showed a minimal association with the continuous discrepancy score ($\beta = -0.02$; 95% CI, −0.04 to 0.00). Each one-level increase in training level was associated with lower odds of overestimation (OR = 0.81; 95% CI, 0.76 to 0.85) and large overestimation (OR = 0.68; 95% CI, 0.61 to 0.76), and with slightly higher odds of underestimation (OR = 1.06; 95% CI, 1.01 to 1.12).

Advanced/non-routine cases likewise showed a minimal association with the continuous discrepancy score ($\beta = -0.02$; 95% CI, −0.04 to 0.00) but were associated with lower odds of within-one-point agreement (OR = 0.84; 95% CI, 0.79 to 0.90) and higher odds of large underestimation (OR = 1.18; 95% CI, 1.10 to 1.27) than basic/routine cases.

After adjustment, records involving male residents showed less negative discrepancy scores ($\beta = 0.13$; 95% CI, 0.01 to 0.25), higher odds of overestimation (OR = 1.39; 95% CI, 1.05 to 1.83), and higher odds of large overestimation (OR = 2.17; 95% CI, 1.34 to 3.51) than records involving female residents. Similarly, records involving male faculty raters showed less negative discrepancy scores ($\beta = 0.20$; 95% CI, 0.06 to 0.34) and lower odds of underestimation (OR = 0.72; 95% CI, 0.54 to 0.98) and large underestimation (OR = 0.63; 95% CI, 0.43 to 0.94) than records involving female faculty raters.

Sensitivity analyses

When training level was modeled categorically, associations were directionally consistent with the ordinal specification: compared with R1, later training levels generally had lower odds of overestimation (lowest for R5: OR = 0.43; 95% CI, 0.34 to 0.55) and higher odds of underestimation (highest for R4: OR = 1.43; 95% CI, 1.21 to 1.68), although the pattern for the continuous discrepancy score was not strictly monotonic (Additional file 1: Supplementary Table S3).

Models using the 12 individual EPA titles instead of the four investigator-defined EPA groups yielded associations similar in direction and magnitude to the primary model; for example, training level remained associated with lower odds of overestimation (OR = 0.81; 95% CI, 0.76 to 0.85), advanced/non-routine cases with higher odds of large underestimation (OR = 1.18; 95% CI, 1.10 to 1.26), and male residents with higher odds of large overestimation (OR = 2.17; 95% CI, 1.34 to 3.50).

Exploratory gender-pairing analysis

In the exploratory gender-pairing model (Additional file 1: Supplementary Table S4), all pairings showed more negative adjusted discrepancy scores than the male resident–male faculty rater reference group: $\beta = -0.31$ (95% CI, −0.50 to −0.13) for female resident–female faculty rater pairs, $\beta = -0.21$ (95% CI, −0.36 to −0.07) for male resident–female faculty rater pairs, and $\beta = -0.14$ (95% CI, −0.26 to −0.02) for female resident–male faculty rater pairs.

Discussion

In this national analysis of 47,346 paired resident and faculty ratings, three principal findings emerged. First, resident self-scores were systematically but modestly lower than faculty scores, with a mean discrepancy of −0.6 points and underestimation five times as common as overestimation (48.5% vs. 9.7%). Second, most paired scores (80.2%) were within one point. Third, the direction of discrepancy varied in a patterned way with training level and case complexity, as overestimation declined from R1 to R5 and large underestimation was more frequent in advanced/non-routine cases. The dominant pattern was therefore not severe resident–faculty disagreement but modest, consistent, conservative resident self-rating, suggesting that discrepancy in this WBA system may carry information about assessment perspectives, confidence, and feedback needs rather than simply indicating inaccurate self-assessment.

The observed alignment should be interpreted in light of the platform workflow. Faculty raters could view resident self-scores when completing their assessments; therefore, the ratings were not generated independently. Previous research indicates that prior information about a learner can influence subsequent faculty ratings through assimilation or other context effects [26, 27]. Some of the observed alignment may consequently reflect anchoring or deliberate consideration of the resident's self-assessment during scoring and feedback. The findings therefore describe resident–faculty score differences within the real-world, non-blinded E-MyWay assessment process rather than agreement between independently generated ratings.

This pattern speaks to two competing accounts of self-assessment in health professions education. Traditional research comparing isolated self-ratings against external performance measures has questioned the accuracy of clinician and trainee self-assessment [5]. Subsequent conceptual and empirical work has framed self-assessment as an informed, developmental process shaped by internal judgments, external feedback, standards, and contextual cues [6–8]. In the present study, 80.2% of paired scores were within one point, and the mean discrepancy was −0.6, suggesting a modest, directionally consistent difference between resident and faculty assessment perspectives rather than widespread large disagreement. Residents appeared to apply a more conservative scoring frame than faculty raters. However, because faculty scores served as comparators rather than an error-free gold standard, these findings do not establish residents' absolute calibration and should not automatically be interpreted as evidence of resident error.

Training level shaped the direction more than the magnitude of discrepancy. The adjusted association with the continuous discrepancy score was minimal ($\beta = -0.02$ per level), yet each additional training level was associated with approximately one-fifth lower odds of overestimation and one-third lower odds of large overestimation, and faculty scores exceeded resident self-scores at every level. This declining tendency to overestimate should not be read as

deteriorating self-assessment among senior residents; it is at least equally consistent with more conservative self-monitoring, greater awareness of clinical complexity, and rising internal standards as residents approach independent practice [6-8].

Case complexity showed a distinct contextual pattern. Advanced/non-routine cases were associated with lower odds of within-one-point agreement and higher odds of large underestimation than basic/routine cases. Complex encounters require residents to integrate clinical reasoning, prioritization, procedural judgment, and adaptation to unpredictable patient and contextual factors [14, 15], conditions under which residents plausibly experience — and score — greater uncertainty. Faculty raters may simultaneously apply a broader frame that credits recognition of complexity, appropriate help-seeking, patient safety, and stage-appropriate performance [1, 3]. Advanced cases may therefore be the encounters in which paired ratings are most informative and discrepancy-focused feedback most valuable.

In exploratory analyses, records involving male residents showed higher adjusted odds of overestimation and large overestimation, records involving male faculty raters showed less negative discrepancy scores and lower odds of underestimation, and all resident–faculty gender pairings showed more negative adjusted discrepancy scores than male–male dyads. The direction of these associations is consistent with prior literature on gender, confidence, self-assessment, and feedback in medical training [4, 16-18]. These findings are hypothesis-generating only; the constraints on their interpretation are addressed in the limitations below.

Educational implications

These findings support a practical reframing of resident–faculty discrepancy as an educational signal rather than a measurement problem. Because agreement within one point was the norm, encounters outside that range constituted a minority and may be useful to flag for review or feedback: large underestimation — present in 17.7% of records and more frequent in complex cases — may prompt discussion of uncertainty, unrecognized strengths, or overly strict self-judgment, whereas the rarer large overestimation (2.1%) may warrant explicit feedback about performance gaps. Used in this way, paired ratings can help faculty structure feedback around performance standards, confidence, and readiness for autonomy, consistent with formative WBA and programmatic assessment approaches [1, 2].

Strengths and limitations

This study has several strengths. It analyzed a national, specialty-wide dataset of 47,346 paired assessments contributed by 466 residents and 447 faculty raters across all 12 EPAs, permitting encounter-level comparison of the two rating perspectives at a scale rarely available in WBA research. The mixed-effects models accounted for clustering by both resident and faculty rater; multiple complementary discrepancy outcomes were examined rather than a single index; and sensitivity analyses using alternative training-level and EPA specifications yielded associations similar in direction and magnitude, supporting the robustness of the main findings.

Several limitations should be considered. First, the design was observational and used routinely collected data, so no causal inference is intended; unmeasured factors—including faculty–resident familiarity, the amount of direct observation, feedback quality, resident confidence, clinical workload, and local assessment culture—may have influenced scoring patterns. Second, faculty scores served as the comparators rather than an error-free gold standard, so discrepancies reflect differences between two assessment perspectives rather than resident error. Moreover, faculty raters could view resident self-scores when completing their assessments; therefore, the ratings were not generated independently. This visibility may have increased apparent agreement or reduced observed discrepancies through anchoring or deliberate educational consideration of residents' self-assessments, although its direction and extent cannot be determined from the available data. Third, the gender-related findings must not be interpreted beyond their exploratory status: gender was available only as a recorded binary variable, and the observed differences could reflect resident self-rating behavior, faculty rating behavior, case assignment, specialty culture, dyadic interpersonal dynamics, or other unmeasured contextual factors; they provide no evidence of causal gender effects or of bias. Fourth, the four EPA groups were investigator-defined, although sensitivity analyses using the 12 individual EPA titles produced similar results. Fifth, the study examined numeric global scores from a single national otolaryngology platform and did not analyze resident reflections, faculty narrative feedback, or entrustment-supervision ratings; this limits generalizability to other specialties and assessment systems and leaves the reasoning behind score differences unexamined.

Because the gender-related findings were exploratory, future mixed-methods research should investigate how resident gender, faculty gender, dyadic interaction, assessment context, and local training culture shape WBA scoring and feedback. Intervention studies could also test whether structured discrepancy-based feedback improves residents' self-assessment calibration, confidence, and self-regulated learning.

Conclusion

In this national otolaryngology WBA dataset of 47,346 paired assessments, residents rated their performance modestly lower than faculty raters (mean discrepancy, –0.6 points), yet 80.2% of paired scores agreed within one point. Underestimation was substantially more common than overestimation, the tendency to overestimate declined with training level, and advanced/non-routine cases showed lower agreement and more frequent large underestimation. These patterns characterize resident–faculty discrepancy as a structured feature of workplace-based assessment rather than noise or resident error. Within these bounds, discrepancy information may serve as an educational signal for feedback, self-assessment calibration, and learner development, while the exploratory gender-related findings await confirmation in future research.

Abbreviations

CBME

Competency-Based Medical Education
EPA
Entrustable Professional Activity
WBA
workplace-based assessments

Declarations

Clinical trial number

not applicable.

Ethics approval and consent to participate

The Institutional Review Board of Cardinal Tien Hospital (CTH-114-2-1-039) approved this study. All participants provided electronic informed consent before their initial E-MyWay platform access. This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Funding

This study was supported by the National Science and Technology Council of the Republic of China (Taiwan) under grants NSTC 110-2511-H-567-001-MY2, NSTC 112-2410-H-567-001-MY3, NSTC 115-2410-H-567-001-MY3 and, in part, funded by Cardinal Tien Hospital under grant CTH113A-2204 and CTH114A-2208. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Author Contribution

Conceptualization: CLH, YJC and JWC.

Data curation: ODL and YHW.

Methodology/formal analysis/validation: ODL, YJC, YHW and MCC.

Project administration: CLH, CHL and JWC.

Funding acquisition: CLH and JWC.

Visualization: ODL and YHW.

Writing – original draft: ODL, CLH and JWC.

Writing – review & editing: ODL, YJC, CLH, YHW, CHL, MCC and JWC.

Acknowledgement

We thank the Taiwan Society of Otorhinolaryngology-Head and Neck Surgery and all its faculties and resident physicians, for utilizing the Joint Commission of Taiwan's E-MyWay platform. We also thank the information technology team of Dalin Tzu Chi Hospital for their support with the platform. Additionally, we are grateful for the administrative assistance provided by Chiu-Ping Wang, Shu-Hwei Fan, Uan-Shr Jan, and Wan-Ning Luo in this project. They received no additional compensation for their contributions.

Data Availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide 31. *Med Teach*. 2007;29(9):855–71.
2. van der Vleuten CP, Schuwirth LW, Driessen EW, Dijkstra J, Tigelaar D, Baartman LK, van Tartwijk J. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205–14.
3. Cate OT. A primer on entrustable professional activities. *Korean J Med Educ*. 2018;30(1):1–10.
4. Blanch DC, Hall JA, Roter DL, Frankel RM. Medical student gender and issues of confidence. *Patient Educ Couns*. 2008;72(3):374–81.
5. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA*. 2006;296(9):1094–102.
6. Eva KW, Regehr G. Self-assessment in the health professions: a reformulation and research agenda. *Acad Med*. 2005;80(10 Suppl):S46–54.
7. Sargeant J, Armson H, Chesluk B, Dornan T, Eva K, Holmboe E, Lockyer J, Loney E, Mann K, van der Vleuten C. The processes and dimensions of informed self-assessment: a conceptual model. *Acad Med*. 2010;85(7):1212–20.
8. Sargeant J, Eva KW, Armson H, Chesluk B, Dornan T, Holmboe E, Lockyer JM, Loney E, Mann KV, van der Vleuten CP. Features of assessment learners use to make informed self-assessments of clinical performance. *Med Educ*. 2011;45(6):636–47.
9. Kendrick DE, Clark MJ, Fischer I, Bohnen JD, Kim GJ, George BC. The reliability of resident self-evaluation of operative performance. *Am J Surg*. 2021;222(2):341–6.
10. Peyre SE, MacDonald H, Al-Marayati L, Templeman C, Muderspach LI. Resident self-assessment versus faculty assessment of laparoscopic technical skills using a global rating scale. *Int J Med Educ*. 2010;1:37–41.
11. Lee JR, Ross S. A Comparison of Resident-Completed and Preceptor-Completed Formative Workplace-Based Assessments in a Competency-Based Medical Education Program. *Fam Med*. 2022;54(8):599–605.
12. Xu X, Ruan X, Yu C, Yu X, Quan X, Li X, Yuan T, Xia D, Zhang Y, Pei L. Differences between resident self-assessments and faculty- assessments on Anesthesiology Milestones and associated factors: a cross-sectional study. *BMC Med Educ*. 2024;24(1):551.
13. Guo FC, Chen YT, Hsu WC, Wang PC, Chen M, Chen JW. EMYWAY Workplace-Based Entrustable Professional Activities Assessments in Otolaryngology Residency Training: A Nationwide Experience. *Otolaryngol Head Neck Surg*. 2025;172(4):1242–53.
14. Cooke S, Lemay JF. Transforming Medical Assessment: Integrating Uncertainty Into the Evaluation of Clinical Reasoning in Medical Education. *Acad Med*. 2017;92(6):746–51.
15. Ilgen JS, Eva KW, de Bruin A, Cook DA, Regehr G. Comfort with uncertainty: reframing our conceptions of how clinicians navigate complex clinical situations. *Adv Health Sci Educ Theory Pract*. 2019;24(4):797–809.
16. 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.
17. Klein R, Julian KA, Snyder ED, Koch J, Ufere NN, Volerman A, Vandenberg AE, Schaeffer S, Palamara K. From the Gender Equity in Medicine (GEM) workgroup: Gender Bias in Resident Assessment in Graduate Medical Education: Review of the Literature. *J Gen Intern Med*. 2019;34(5):712–9.
18. Mueller AS, Jenkins TM, Osborne M, Dayal A, O'Connor DM, Arora VM. Gender Differences in Attending Physicians' Feedback to Residents: A Qualitative Analysis. *J Grad Med Educ*. 2017;9(5):577–85.
19. Chen JW, Tu HL, Chang CH, Hsu WC, Wang PC, Liao CH, Chen M. Automated Evaluation of Reflection and Feedback Quality in Workplace-Based Assessments by Using Natural Language Processing: Cross-Sectional Competency-Based Medical Education Study. *JMIR Med Educ*. 2025;11:e81718.
20. Chen JW, Yu RB, Lin KN, Ono Y, Yu XB, Hsu WC, Wang PC, Liao LJ, Liao CH, Chen M. Learning Analytics of a National Entrustable Professional Activities Platform: Cross-Sectional Study of System-Level Constraints on Advanced Entrustment in Competency-Based Medical Education. *JMIR Med Educ*. 2026;12:e95066.
21. Lin CP, Hsu WC, Wang GL, Su HC, Chung HC, Hung SH, Chen M, Chen JW, Wang PC. Factors associated with attainment of ad hoc entrustability among Taiwan otolaryngology resident physicians: a nationwide cross-sectional study. *BMC Med Educ*. 2025;25(1):1135.
22. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, Initiative S. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet*. 2007;370(9596):1453–7.
23. R Core Team. R: A language and environment for statistical computing. In., Version 4.5.3 edn. Vienna, Austria: R Foundation for Statistical Computing; 2026.
24. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw*. 2015;67(1):1–48.
25. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest Package: Tests in Linear Mixed Effects Models. *J Stat Softw*. 2017;82(13):1–26.
26. Humphrey-Murto S, LeBlanc A, Touchie C, Pugh D, Wood TJ, Cowley L, et al. The influence of prior performance information on ratings of current performance and implications for learner handover: a scoping review. *Acad Med*. 2019;94(7):1050–7.
27. Shaw T, Wood TJ, Touchie C, Pugh D, Humphrey-Murto SM. How biased are you? The effect of prior performance information on attending physician ratings and implications for learner handover. *Adv Health Sci Educ Theory Pract*. 2021;26(1):199–214.

Tables

Table 1
Assessment-record characteristics of the final analytic sample

Characteristic	Category	WBA records, n (%)
Training level	R1	10048 (21.2%)
	R2	9721 (20.5%)
	R3	9778 (20.7%)
	R4	9486 (20.0%)
	R5	8313 (17.6%)
EPA group	Early/basic	16571 (35.0%)
	Core disease-management	15461 (32.7%)
	Delayed/advanced	13828 (29.2%)
	Presentation/communication	1486 (3.1%)
Case complexity	Basic/routine	32078 (67.8%)
	Advanced/non-routine	15266 (32.2%)
	Missing	2 (0.0%)
Resident gender	Female	13414 (28.3%)
	Male	33932 (71.7%)
Faculty rater gender	Female	7042 (14.9%)
	Male	40304 (85.1%)
Resident–faculty gender pairing	Female resident–female faculty rater	2179 (4.6%)
	Female resident–male faculty rater	11235 (23.7%)
	Male resident–female faculty rater	4863 (10.3%)
	Male resident–male faculty rater	29069 (61.4%)
Note. Percentages are based on 47346 workplace-based assessment (WBA) records. EPA groups were investigator-defined task-context categories: early/basic = EPA01–EPA05; core disease-management = EPA07–EPA09; delayed/advanced = EPA06, EPA10, and EPA11; presentation/communication = EPA12.		

Table 2
Resident–faculty score discrepancy outcomes overall and by assessment characteristics

Group	Subgroup	n	Resident self-score, mean $\pm$ SD	Faculty score, mean $\pm$ SD	Discrepancy score, mean $\pm$ SD	Overestimation	Exact agreement	Underestimation	Within one-point agree
Overall	Overall	47346	7.4 $\pm$ 1.4	8.0 $\pm$ 1.3	-0.6 $\pm$ 1.2	9.7%	41.8%	48.5%	80.2%
Training level	R1	10048	7.0 $\pm$ 1.6	7.5 $\pm$ 1.4	-0.5 $\pm$ 1.2	13.9%	42.3%	43.8%	80.3%
	R2	9721	7.2 $\pm$ 1.4	7.8 $\pm$ 1.4	-0.6 $\pm$ 1.2	10.5%	44.0%	45.5%	80.7%
	R3	9778	7.4 $\pm$ 1.3	8.1 $\pm$ 1.2	-0.7 $\pm$ 1.1	9.1%	40.8%	50.1%	79.5%
	R4	9486	7.6 $\pm$ 1.3	8.2 $\pm$ 1.2	-0.7 $\pm$ 1.1	7.7%	40.5%	51.7%	80.0%
	R5	8313	7.8 $\pm$ 1.3	8.5 $\pm$ 1.2	-0.7 $\pm$ 1.0	7.0%	40.9%	52.1%	80.5%
Case complexity	Basic/routine	32078	7.4 $\pm$ 1.4	8.0 $\pm$ 1.3	-0.6 $\pm$ 1.1	9.7%	44.0%	46.3%	82.4%
	Advanced/non-routine	15266	7.3 $\pm$ 1.4	8.1 $\pm$ 1.4	-0.8 $\pm$ 1.2	9.8%	37.1%	53.2%	75.5%
EPA group	Early/basic	16571	7.5 $\pm$ 1.4	8.2 $\pm$ 1.3	-0.6 $\pm$ 1.2	9.7%	42.3%	48.0%	80.1%
	Core disease-management	15461	7.3 $\pm$ 1.5	7.9 $\pm$ 1.3	-0.6 $\pm$ 1.2	10.2%	41.8%	48.1%	80.0%
	Delayed/advanced	13828	7.3 $\pm$ 1.4	7.9 $\pm$ 1.3	-0.6 $\pm$ 1.1	9.6%	41.4%	49.0%	80.8%
	Presentation/communication	1486	7.4 $\pm$ 1.4	8.1 $\pm$ 1.3	-0.7 $\pm$ 1.1	8.0%	38.2%	53.8%	77.1%
Resident gender	Female	13414	7.4 $\pm$ 1.4	8.1 $\pm$ 1.2	-0.7 $\pm$ 1.2	7.6%	40.1%	52.4%	77.5%
	Male	33932	7.4 $\pm$ 1.5	8.0 $\pm$ 1.4	-0.6 $\pm$ 1.1	10.6%	42.4%	47.0%	81.3%
Faculty rater gender	Female	7042	7.4 $\pm$ 1.5	8.1 $\pm$ 1.3	-0.7 $\pm$ 1.2	7.9%	40.4%	51.6%	79.1%
	Male	40304	7.4 $\pm$ 1.4	8.0 $\pm$ 1.3	-0.6 $\pm$ 1.1	10.1%	42.0%	47.9%	80.4%

Note. Discrepancy score was calculated as resident self-score minus faculty global performance score; negative values indicate lower resident self-scores relative to faculty scores. Overestimation/underestimation indicate resident self-scores higher/lower than faculty scores. Exact agreement indicates identical scores. Within-one-point agreement indicates $|\text{discrepancy}| \leq 1$. Two records with missing case-complexity data were excluded from the case-complexity subgroup rows.

Table 3
Adjusted mixed-effects models for resident–faculty score discrepancy outcomes

Predictor	Discrepancy score, β (95% CI)	Overestimation, OR (95% CI)	Underestimation, OR (95% CI)	Within-one-point agreement, OR (95% CI)	Large overestimation, OR (95% CI)	Large underestimation, OR (95% CI)
Training level, per one-level increase	-0.02 (-0.04, 0.00)	0.81 (0.76, 0.85)	1.06 (1.01, 1.12)	1.05 (0.99, 1.11)	0.68 (0.61, 0.76)	1.00 (0.94, 1.07)
Advanced/non-routine case	-0.02 (-0.04, 0.00)	1.20 (1.11, 1.30)	1.06 (1.01, 1.12)	0.84 (0.79, 0.90)	1.16 (0.98, 1.37)	1.18 (1.10, 1.27)
Male resident	0.13 (0.01, 0.25)	1.39 (1.05, 1.83)	0.82 (0.62, 1.07)	1.09 (0.83, 1.42)	2.17 (1.34, 3.51)	0.78 (0.55, 1.09)
Male faculty rater	0.20 (0.06, 0.34)	1.36 (0.99, 1.89)	0.72 (0.54, 0.98)	1.38 (0.99, 1.93)	1.49 (0.85, 2.63)	0.63 (0.43, 0.94)

Note. Analytic sample for all adjusted models: $n = 47344$. Values are β coefficients for the continuous discrepancy-score model and odds ratios for binary outcomes, with 95% confidence intervals in parentheses. Models included random intercepts for resident and faculty rater. Fixed effects included the predictors shown in the table plus EPA group, clinical setting, hospital level, region, and assessment period. Table 3 reports selected secondary contextual predictors and exploratory gender-related variables.

Figures

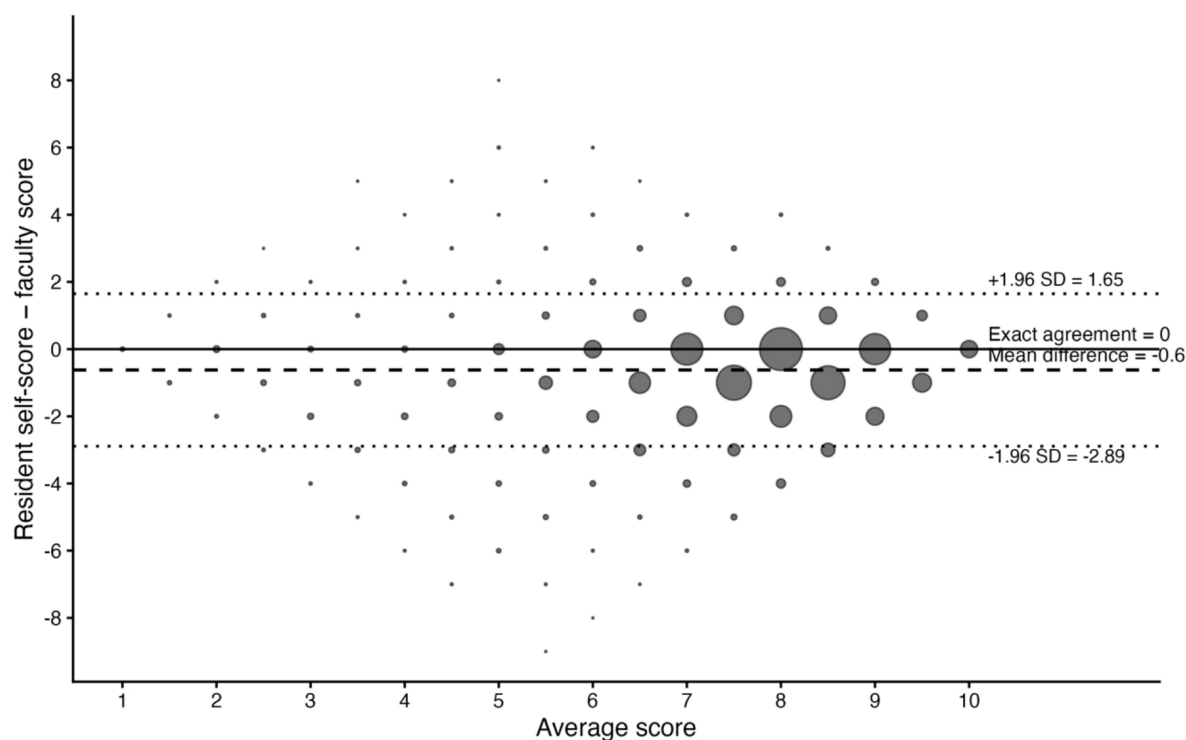


Figure 1

Bland–Altman plot of resident–faculty score discrepancy. The difference between the paired scores, calculated as resident self-score minus faculty global performance score, is plotted against their mean for each workplace-based assessment record. Bubble size represents the number of records at each coordinate. The solid horizontal line at 0 indicates exact agreement, the dashed line indicates the mean discrepancy (-0.6), and the dotted lines indicate the 95% limits of agreement (lower, -2.89 ; upper, 1.65). Negative values indicate that resident self-scores were lower than faculty scores.

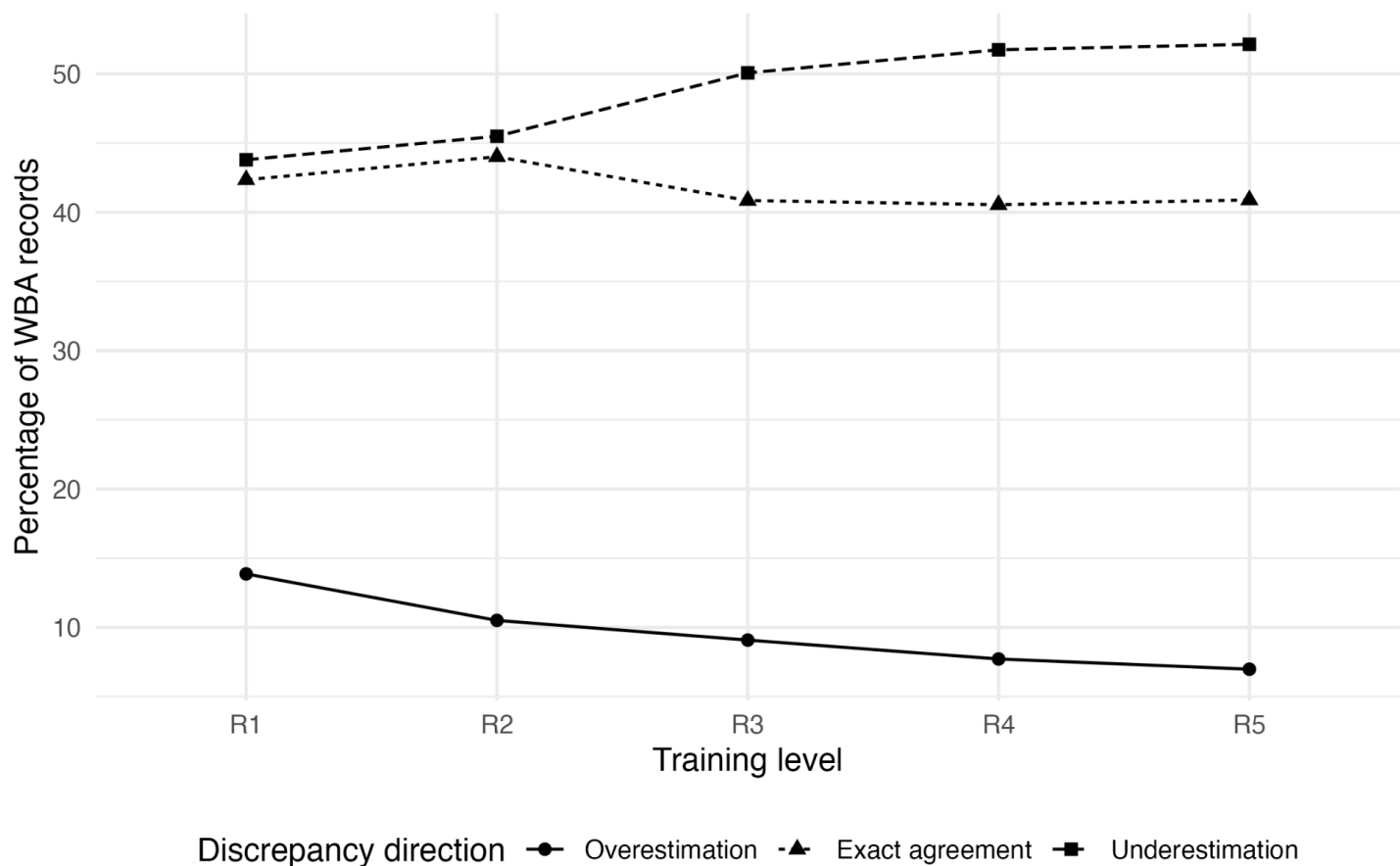


Figure 2

Direction of resident–faculty discrepancy by training level. Lines with points show the percentage of WBA records classified as overestimation, exact agreement, or underestimation within each training level. Overestimation was defined as a resident self-score higher than the faculty score; exact agreement as identical resident self-score and faculty score; and underestimation as a resident self-score lower than the faculty score.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterial.docx](#)