Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prior Simulation Training and Gender Differences in Instructor Self-Assessment of Debriefing

Jordi Bañeras

Jordi.Baneras@uab.cat

Vall d'Hebron Hospital Universitari

Ainoa Ares

Universitat Autònoma de Barcelona

Francisco Javier Tornero

Universitat Autònoma de Barcelona

Albert Selva

Universitat Autònoma de Barcelona

Roser Solans

Universitat Autònoma de Barcelona

Laura Gisbert

Universitat Autònoma de Barcelona

Nieves Barragan

Universitat Autònoma de Barcelona

Nayana Joshi

Universitat Autònoma de Barcelona

Eloi Espín

Universitat Autònoma de Barcelona

Sonia Cañadas

Universitat Autònoma de Barcelona

Antonio Gil-Moreno

Universitat Autònoma de Barcelona

Miguel Angel Gonzalez-Posada

Universitat Autònoma de Barcelona

Laura Tortosa

Universitat Autònoma de Barcelona

Gemma Cervero

Universitat Autònoma de Barcelona

Marta Sanchez

Vall d'Hebron Hospital Universitari

Monica Rodriguez-Carballeira

Vall d'Hebron Hospital Universitari

Miriam de Nadal

Universitat Autònoma de Barcelona

Research Article

Keywords: debriefing, faculty development, gender, self-assessment, simulation-based education, DASH

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11398317/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

Faculty development is intended to strengthen debriefing practice, yet evidence is limited on whether prior simulation training is associated with instructors' own Debriefing Assessment for Simulation in Healthcare (DASH) ratings or whether those ratings vary by gender.

Objective

To compare instructor-reported DASH scores by prior simulation training and gender and to examine training intensity, training-by-gender interaction, and scenario-specific heterogeneity.

Methods

Repeated cross-sectional analysis of 721 consented DASH Instructor Version-Short Form questionnaires collected during 2024–2025 and 2025–2026 across 22 recurring final-year medical student scenarios delivered in both academic years. Linear models included harmonized scenario and academic-year fixed effects with cluster-robust standard errors. Training models adjusted for gender; gender models adjusted for training. Item-level P values were Holm corrected. Top-box and interaction analyses were prespecified secondary analyses.

Results

The overall global score was 6.49 ± 0.55. Any documented prior training was not associated with the global score (adjusted mean difference, -0.03; 95% CI, -0.13 to 0.06; P = 0.483) or any item. Women rated themselves lower than men globally (-0.15; 95% CI, -0.23 to -0.08; P < 0.001) and on all six items after correction; differences were greatest for organized structure (-0.23) and in-depth discussion (-0.21). Women had lower odds of assigning 7 to all six items (OR, 0.40; 95% CI, 0.27 to 0.59). Instructor-course ratings were lower than basic/other training (-0.20; 95% CI, -0.29 to -0.10), and the gender difference was concentrated in the instructor-course group (interaction P < 0.001). Gender effects did not differ significantly across scenarios (P = 0.155).

Conclusions

Prior training was not associated with higher instructor-reported DASH scores, whereas women reported lower scores across all six domains. The concentration of lower ratings among formally trained women suggests that training context and gender may shape how instructors apply DASH criteria to their own

debriefings. The DASH Instructor Version (DASH-IV) may therefore be most informative as a structured reflective and faculty-development measure rather than as a standalone indicator of competence.

Introduction

Simulation programs increasingly invest in faculty development because the educational value of debriefing depends on facilitators' ability to create an engaging climate, organize reflection, explore performance gaps, and translate discussion into future action. The 2025 Healthcare Simulation Standards of Best Practice therefore position facilitator preparation and continuing development as core program responsibilities.[1] Yet debriefing methods and faculty-development pathways remain highly variable across institutions.[2]

Recent reviews show that simulation faculty development differs markedly in duration, content, deliberate practice, coaching, and outcome measurement.[3] A 2026 systematic review found generally favorable effects of debriefing-focused interventions on facilitator skill, confidence, and preparedness, while emphasizing heterogeneous designs and limited learner-level evidence.[4] Individual courses and peer-coaching initiatives have improved assessed debriefing behaviors,[5,6] but whether prior training is associated with higher ratings during routine curricular delivery remains uncertain.

The Debriefing Assessment for Simulation in Healthcare (DASH) provides a behaviorally anchored framework for examining six domains of debriefing practice.[7] A Spanish version has undergone linguistic and cultural validation,[8] and the Instructor Version-Short Form enables debriefers to review their own performance immediately after a session. However, ratings vary according to who completes the instrument; when student, trained-rater, and instructor versions were applied to the same teledebriefings, instructors assigned lower scores than students and raters.[9] This is consistent with broader evidence that professional self-assessment is a metacognitive judgment shaped by available standards, feedback, and insight, rather than a direct proxy for observed competence.[10,11]

This distinction is especially relevant when comparing instructor groups. Studies in medical education have reported lower self-ratings among women despite similar objective performance,[12] and longitudinal evidence suggests that gender-related differences in self-assessment persist after accounting for knowledge and academic progression.[13] Whether analogous patterns occur among simulation faculty, whether they concern specific DASH domains, and whether they vary according to the intensity of prior simulation training have not been established.

Accordingly, our primary objectives were to determine whether instructor-reported DASH scores differed between sessions evaluated by instructors with and without prior simulation training and to compare global and item-level scores between women and men. Strategic secondary objectives were to examine training intensity, test whether the association between gender and self-ratings differed by training level, evaluate scenario-specific heterogeneity, and assess the robustness of findings using top-box outcomes.

Methods

Study Design and Setting

We conducted a repeated cross-sectional observational study of routinely collected instructor self-assessments from the Clinical Simulation subject in the final year of the medical curriculum at the Teaching Unit of Vall d'Hebron University Hospital, Faculty of Medicine, Universitat Autònoma de Barcelona, Spain. The present analysis included the 2024–2025 and 2025–2026 academic years, during which instructor gender and prior simulation training were recorded. Data from 2023–2024 were not eligible for the primary objectives because these exposure variables had not yet been collected.

Participants and Simulation Program

Faculty from seven participating departments facilitated 22 clinical scenarios in each academic year: Family and Community Medicine (3 scenarios), Gynecology and Obstetrics (2), Medicine (6), Orthopedic Surgery and Traumatology-Anesthesiology (2), Pediatrics (3), Psychiatry (4), and Surgery (2). This corresponded to 44 scenario-year implementations across the two-year analytic period. The psychiatry scenario addressing a first psychotic episode was recorded under slightly different labels in the two academic years and was harmonized before analysis. Each session included a briefing, a clinical scenario, and a structured plus-delta debriefing. All instructors who completed the post-session quality questionnaire were eligible. Participation in the questionnaire was voluntary, and responses without consent for research use were excluded.

DASH Measure and Outcomes

Immediately after each session, instructors completed the Spanish DASH Instructor Version-Short Form (DASH-IV) using a QR-code questionnaire.[8] The instrument contains six corresponding 7-point items: establishing an engaging learning environment; maintaining an engaging context; structuring the debriefing in an organized manner; provoking in-depth discussion and reflection; identifying strengths, performance gaps, and their causes; and helping participants improve or sustain good performance. The primary global outcome was the arithmetic mean of the six items. Item-level scores were secondary outcomes. Top-box outcomes were defined as a score of 7 for each item and a score of 7 on all six items. Internal consistency was summarized using Cronbach alpha.

Exposure Definitions

Prior training was obtained from a structured free-text/multiple-response field and coded before outcome modeling. Responses explicitly stating no prior simulation education were classified as no training. Documented introductory, online, workshop-based, or other basic simulation preparation was classified as basic/other prior training. Completion of a formal simulation instructor course was

classified as instructor-course training. Ambiguous or incomplete responses were retained for descriptive reporting but excluded from training-effect models. The prespecified primary training exposure compared any clearly documented prior training with no training. Training intensity (none, basic/other, instructor course) was analyzed secondarily. Gender was self-reported as woman, man, nonbinary, or prefer not to disclose. Because the latter two categories contained seven responses in total, binary gender comparisons included women and men only; excluded categories remain represented in the overall descriptive cohort.

Statistical Analysis

Continuous outcomes are reported as mean and standard deviation and categorical outcomes as number and percentage. Primary associations were estimated using linear regression with clinical-scenario fixed effects and academic-year fixed effects. The training models were additionally adjusted for gender, and the gender models were adjusted for training category. Cluster-robust standard errors accounted for nonindependence within academic year, session date, and clinical scenario. The global DASH score was the prespecified primary outcome. For the six item-level analyses, two-sided P values were adjusted using the Holm method. Adjusted mean differences are reported with 95% confidence intervals. Binomial generalized linear models with the same covariate structure evaluated top-box outcomes and are reported as odds ratios. Secondary models compared the three training levels, tested training-level-by-gender interactions, and estimated scenario-specific gender and training effects. Scenario-specific P values were controlled using the Benjamini-Hochberg false-discovery-rate procedure, and joint interaction tests assessed heterogeneity across scenarios. Because questionnaires were anonymous, responses could not be linked to individual instructors; therefore, instructor-level random effects could not be fitted. No a priori sample-size calculation was performed because all eligible routine evaluations were included. Analyses were conducted using SPSS V 21. Statistical significance was set at $P < 0.05$.

Ethics

The study was conducted in accordance with the Declaration of Helsinki. The Research Ethics Committee of the "Universitat Autònoma de Barcelona" reviewed the institutional protocol (reference number SCER141) and confirmed exemption from full ethics review under university policy. Participation was voluntary, and all participating instructors provided informed consent to participate and for their anonymized data to be analyzed and published through a consent item included in the questionnaire; responses without consent were excluded. No directly identifying information was collected.

Results

The database contained 1,092 instructor evaluations across three academic years. The 368 evaluations from 2023–2024 were excluded because gender and training were not collected. Of 724 evaluations submitted during 2024–2025 and 2025–2026, three lacked consent, leaving 721 questionnaires (99.6% of submitted evaluations) for analysis. The curriculum comprised 22 recurring clinical scenarios in each

academic year, corresponding to 44 scenario-year implementations. There were 385 evaluations in 2024–2025 and 336 in 2025–2026. Medicine contributed 166 evaluations (23.0%), Psychiatry 145 (20.1%), Family and Community Medicine 125 (17.3%), Pediatrics 117 (16.2%), Orthopedic Surgery and Traumatology-Anesthesiology 73 (10.1%), Gynecology and Obstetrics 61 (8.5%), and Surgery 34 (4.7%). Disciplines with larger scenario portfolios generally contributed more questionnaires, although contributions were not strictly proportional to scenario count. These percentages describe the distribution of analyzed questionnaires; a true response rate could not be calculated because the number of eligible instructor-session assessments was not recorded.

The overall global DASH score was 6.49 ± 0.55. Item means ranged from 6.44 ± 0.70 for provoking in-depth discussion to 6.51 ± 0.60 for maintaining an engaging context. Cronbach alpha was 0.937. Between 54.8% and 56.2% of item responses were scored 7, and 312 of 721 questionnaires (43.3%) assigned a score of 7 to all six items, indicating substantial concentration at the upper end of the scale (Table 1).

Table 1
Study Sample and Descriptive DASH Instructor Self-Assessments

Characteristic	n (%) or mean ± SD	Notes
Eligible evaluations, 2024–2026	724	Exposure variables collected
Consented analytic cohort	721 (99.6%)	22 recurring scenarios; 44 scenario-year implementations
Academic year 2024–2025	385 (53.4%)	
Academic year 2025–2026	336 (46.6%)	
Gender: man	377 (52.3%)	
Gender: woman	337 (46.7%)	
Gender: nonbinary/prefer not to disclose	7 (1.0%)	Retained descriptively
Training: none	195 (27.0%)	Explicitly reported no training
Training: basic/other prior training	236 (32.7%)	Introductory, online, workshop, or other
Training: instructor course	265 (36.8%)	Formal simulation instructor course
Training: ambiguous/incomplete	25 (3.5%)	Excluded from training-effect models
Global DASH score	6.49 ± 0.55	Range, 1–7
Item 1: engaging environment	6.50 ± 0.59	Top-box 55.5%
Item 2: engaging context	6.51 ± 0.60	Top-box 56.2%
Item 3: organized structure	6.47 ± 0.65	Top-box 54.8%
Item 4: in-depth discussion	6.44 ± 0.70	Top-box 55.1%
Item 5: strengths/gaps and why	6.50 ± 0.62	Top-box 55.8%
Item 6: improvement guidance	6.49 ± 0.60	Top-box 54.8%
All six items scored 7	312 (43.3%)	
Cronbach alpha	0.937	Internal consistency

Prior training was classified as none in 195 evaluations, basic/other in 236, instructor-course training in 265, and ambiguous/incomplete in 25. For the primary adjusted training analysis, ambiguous training responses and seven responses outside the binary gender categories were excluded, leaving 689 evaluations. In unadjusted descriptive analysis, the global score was 6.46 ± 0.54 without training and 6.50 ± 0.55 with any documented prior training. After adjustment, any prior training was not associated with the global score (adjusted mean difference, -0.03; 95% CI, -0.13 to 0.06; P = 0.483) or with any

individual item; all Holm-adjusted P values were at least 0.998 (Table 2 and Fig. 1A). The odds of assigning 7 to all six items were also similar (OR, 0.86; 95% CI, 0.51 to 1.47; P = 0.589).

Table 2
Adjusted Differences in Global and Item-Level DASH Scores by Prior Training and Gender

Outcome	Training difference*	95% CI	P	Gender difference†	95% CI	P‡
Global DASH score	-0.033	-0.127 to + 0.060	0.483	-0.155	-0.228 to -0.082	< 0.001
Item 1	-0.045	-0.144 to + 0.054	1.000	-0.113	-0.196 to -0.030	0.017
Item 2	-0.075	-0.182 to + 0.031	0.998	-0.121	-0.207 to -0.035	0.017
Item 3	-0.014	-0.131 to + 0.104	1.000	-0.229	-0.316 to -0.141	< 0.001
Item 4	+ 0.003	-0.123 to + 0.129	1.000	-0.213	-0.313 to -0.113	< 0.001
Item 5	-0.052	-0.160 to + 0.057	1.000	-0.144	-0.230 to -0.058	0.004
Item 6	-0.017	-0.117 to + 0.083	1.000	-0.110	-0.191 to -0.029	0.017

* Any documented prior training minus no training; adjusted for gender, academic year, and clinical scenario. † Women minus men; adjusted for training category, academic year, and clinical scenario. ‡ Global-score P values are unadjusted; item-level P values are Holm adjusted. Cluster-robust 95% confidence intervals.

Gender comparisons included 377 evaluations from men and 337 from women (n = 714). The unadjusted global mean was 6.58 ± 0.53 for men and 6.39 ± 0.55 for women. After adjustment for training, academic year, and scenario, women's global self-ratings were 0.15 points lower than men's (95% CI, -0.23 to -0.08; P < 0.001). Lower ratings among women were present for all six items after Holm correction. The largest differences occurred for organized structure (item 3: -0.23; 95% CI, -0.32 to -0.14) and in-depth discussion (item 4: -0.21; 95% CI, -0.31 to -0.11) (Table 2 and Fig. 1B). Women also had lower adjusted odds of assigning 7 to all six items (OR, 0.40; 95% CI, 0.27 to 0.59; P < 0.001).

In the training-intensity analysis, basic/other preparation did not differ from no training in the global score (adjusted difference, 0.07; 95% CI, -0.04 to 0.18; P = 0.195). Instructor-course evaluations were lower than no-training evaluations (-0.13; 95% CI, -0.23 to -0.02; P = 0.016) and lower than basic/other training (-0.20; 95% CI, -0.29 to -0.10; P < 0.001). This pattern was most consistent when the instructor-course group was compared with basic/other training across the six items (Table 3).

Table 3
Training Intensity and Training-by-Gender Interaction for the Global DASH Score

Contrast	Adjusted difference	95% CI	P
Basic/other training vs no training	+ 0.070	-0.036 to + 0.175	0.195
Instructor course vs no training	-0.125	-0.228 to -0.023	0.016
Instructor course vs basic/other training	-0.195	-0.288 to -0.102	< 0.001
Women vs men: no training	+ 0.011	-0.126 to + 0.149	0.871
Women vs men: basic/other training	-0.101	-0.235 to + 0.032	0.138
Women vs men: instructor course	-0.361	-0.478 to -0.244	< 0.001

The joint test of training-level-by-gender interaction was P = 0.0003. Models adjusted for academic year and clinical scenario with cluster-robust standard errors.

There was evidence that the gender difference varied by training level for the global score (joint interaction P = 0.0003). The adjusted women-minus-men difference was 0.01 points among evaluations reporting no prior training (95% CI, -0.13 to 0.15), -0.10 among those reporting basic/other training (95% CI, -0.23 to 0.03), and − 0.36 among those reporting an instructor course (95% CI, -0.48 to -0.24) (Fig. 2). The interaction was most pronounced for organized structure and in-depth discussion. These analyses were exploratory and were not based on linked longitudinal observations of the same instructors.

The overall gender difference was not demonstrably localized to particular clinical scenarios. The joint test for scenario-by-gender heterogeneity was not significant (P = 0.155), and no scenario-specific gender contrast remained significant after false-discovery-rate correction (Fig. 3; Table, Supplemental Digital Content 1). Training effects varied nominally across scenarios (joint P = 0.003), but several scenario-level strata contained very few or no untrained evaluations, producing unstable estimates; these exploratory results are shown in the Table, Supplemental Digital Content 2.

Discussion

In this two-year study of 721 instructor evaluations from 22 recurring scenarios delivered across 44 scenario-year implementations, the binary comparison of any documented prior simulation training with no training showed no difference in global or item-level DASH self-ratings. In contrast, women reported lower self-ratings than men across all six domains after adjustment for training, academic year, and scenario. The largest differences involved structuring the debriefing and provoking reflective discussion. The gender difference was broadly distributed rather than consistently localized to specific clinical scenarios. A strategic exploratory analysis further indicated that the gender difference was concentrated among evaluations from instructors reporting completion of a formal instructor course.

The absence of a positive binary association between prior training and self-ratings should not be interpreted as evidence that faculty development is ineffective. Reviews published in 2024 and 2026 describe a heterogeneous field in which training programs differ substantially in duration, content, opportunity for deliberate practice, coaching, and outcome measurement.[3,4] Targeted courses and peer-coaching programs have improved externally or learner-rated debriefing behaviors in several settings.[5,6] Our exposure combined diverse prior experiences, was not randomized, and did not capture recency, course duration, faculty experience, or subsequent practice. Moreover, the outcome was instructor self-assessment rather than trained external observation. The null primary finding is therefore more accurately interpreted as an absence of evidence that having any prior training is associated with higher self-reported DASH scores in this routine program.

The lower self-ratings in the instructor-course group are similarly unlikely to establish poorer debriefing performance. Formal faculty development may increase instructors' awareness of the complexity of debriefing, expose them to more demanding behavioral standards, and improve recognition of subtle performance gaps. This response-shift or recalibration explanation is compatible with the broader literature showing that self-assessment is an inferential metacognitive judgment rather than a direct readout of competence.[10,11] It is also consistent with the finding that instructors rated themselves below students and trained raters when different DASH versions were applied to the same debriefings.[9] Determining whether advanced training improves performance while lowering self-ratings requires paired external DASH-Rater Version (DASH-RV) observations before and after training, not cross-sectional self-ratings alone.

The gender findings warrant careful interpretation. Recent studies among medical trainees have found that women may rate their performance lower despite similar examination and clinical assessment results,[12] and longitudinal data suggest persistent gender-related differences in self-assessment accuracy after adjustment for knowledge.[13] Our results extend the question to faculty self-evaluation in simulation, but they do not establish underestimation because no objective benchmark was available. Gender differences in confidence, calibration, response style, professional expectations, or prior feedback could all contribute. The concentration of the difference in items 3 and 4 may be relevant because organizing a complex discussion and eliciting reflective inquiry require nuanced judgments that may be particularly sensitive to personal standards. However, these mechanisms were not measured and should remain hypotheses rather than explanations.

The training-by-gender interaction provides a potentially important, although exploratory, direction for faculty-development research. The near-absence of a gender difference among instructors reporting no training and its progressive emergence across more intensive preparation could indicate differential recalibration after exposure to explicit standards. Alternatively, it may reflect selection into training, differences in clinical specialties, unmeasured experience, or repeated contributions from particular instructors. Because anonymous questionnaires prevented identification of individuals, the analysis cannot determine whether training changed self-ratings within the same instructor. A prospective design

should measure confidence and calibration before training, obtain independent DASH-RV ratings, and follow instructors longitudinally to distinguish learning effects from changes in internal standards. One further, necessarily speculative, interpretation relates to the impostor phenomenon, in which individuals perceive their own performance less favorably than objective indicators would suggest; gender has been associated with impostor experiences among physicians and physicians in training.[19] Because no objective benchmark of debriefing performance was available in this study, this possibility should be regarded as an interpretive hypothesis rather than a conclusion.

Scenario-specific analysis did not identify robust locations for the gender difference after multiplicity correction. This suggests that the overall association was not driven solely by a small number of scenarios or specialties. The nominal heterogeneity in training effects was harder to interpret because some scenarios had almost no untrained evaluations. Future programs should predefine minimum cell sizes or use hierarchical models supported by instructor identifiers and balanced sampling across scenarios.

The study has practical implications for quality assurance. Instructor self-assessment remains useful for structured reflection and identifying personal development goals, but it should form one component of a multisource system. Periodic observation by calibrated DASH-RV raters, learner ratings, video-assisted review, peer coaching, and linkage to instructor training histories would permit triangulation and improve interpretability.[7,9,14] In addition, the heterogeneity of the training received by instructors suggests that programs may benefit from a standardized, structured instructor-preparation pathway completed before facilitating debriefings, which would provide a shared reference standard for applying DASH criteria and could facilitate more consistent debriefing quality. Self-ratings should not be used alone to infer competence, compare genders, or judge the effectiveness of training.

Strengths include the large number of routine evaluations, two academic years, broad clinical coverage, prespecified global and item-level outcomes, adjustment for scenario and academic year, cluster-robust uncertainty estimates, multiplicity correction, top-box sensitivity analyses, and explicit examination of training intensity and interaction. The study also has important limitations. It was conducted at one medical school; training and gender were self-reported; training categories were heterogeneous; and instructor identity, years of experience, professional category, and debriefing duration were unavailable. Repeated questionnaires from the same instructors could not be linked, leaving residual within-instructor dependence. The DASH-IV reflects self-perception, not objective quality, and no DASH-RV or contemporaneous student assessment was available for calibration. The pronounced ceiling effect reduced discrimination. Sparse nonbinary and nondisclosed gender categories precluded meaningful analysis beyond women and men. Finally, observational associations and exploratory interactions cannot support causal conclusions.

Conclusions

Any documented prior simulation training was not associated with higher instructor-reported DASH scores in this routine two-year program. Women reported lower scores than men across all six items, particularly organized structure and reflective discussion, without robust localization to specific scenarios. The concentration of lower ratings among formally trained women suggests that explicit debriefing standards may shape self-evaluation differently across instructor groups. These findings support using DASH-IV as a structured reflection and quality-improvement tool, interpreted in relation to training intensity, gender, and program context rather than as a standalone measure of instructor competence.

Abbreviations

CI: Confidence interval

DASH: Debriefing Assessment for Simulation in Healthcare

DASH-IV: DASH Instructor Version-Short Form

DASH-RV: DASH Rater Version

OR: Odds ratio

SPSS: Statistical Package for the Social Sciences

Declarations

Ethics approval and consent to participate: The study was conducted in accordance with the Declaration of Helsinki. The Research Ethics Committee of the Universitat Autònoma de Barcelona reviewed the institutional protocol (reference number SCER141) and confirmed exemption from full ethics review under university policy. Participation was voluntary, and all participating instructors provided informed consent to participate and for their anonymized data to be analyzed and published through a consent item included in the questionnaire; responses without consent were excluded. No directly identifying information was collected.

Consent for publication: Not applicable.

Availability of data and materials: The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

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

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions: JB led the study; he developed the methodology, conducted the investigation, wrote the original draft of the manuscript, and supervised and validated the work. FJT contributed to data collection and, as the centre's simulation technician, to the preparation of the simulation sessions. AGM, NB, RS and MAGP contributed to writing and editing the final manuscript. MRC and MdN supervised and directed the study and validated the work. LT supervised the data analysis, contributed to the final writing and validation of the manuscript, and managed its submission. GC and MS contributed to writing the final manuscript and provided critical comments. AA, AS, LG, NJ, EE and SC contributed to data collection and delivered the simulation sessions. All authors read and approved the final manuscript.

Acknowledgements: Not applicable.

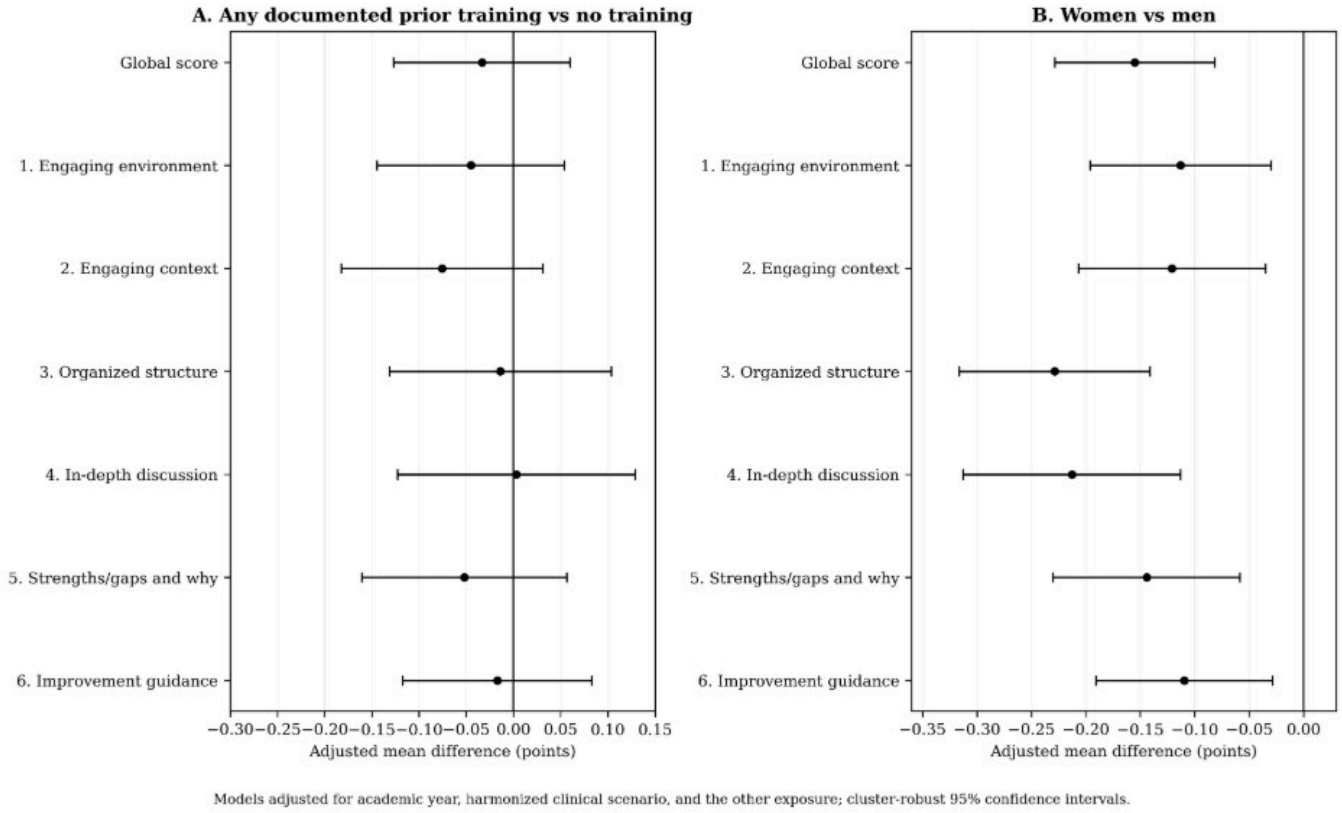
References

1. Standards Committee INACSL, Decker S, Sapp A, Bibin L, Chidume T, Crawford SB, Fayyaz J, Johnson BK, Szydlowski J. Healthcare Simulation Standards of Best Practice®: The Debriefing Process. Clin Simul Nurs. 2025;105:101775.
2. Duff JP, Morse KJ, Seelandt J, Gross IT, Lydston M, Sargeant J, Dieckmann P, Allen JA, Rudolph JW, Kolbe M. Debriefing methods for simulation in healthcare: A systematic review. Simul Healthc. 2024;19(1S):S112–21.
3. Gardner AK, Rodgers DL, Steinert Y, Davis R, Condron C, Peterson DT, Rohra A, Viggers S, Eppich WJ, Reedy G. Mapping the terrain of faculty development for simulation: A scoping review. Simul Healthc. 2024;19(1S):S75–89.
4. Reyes JS, Leal-Giraldo SI, Estupiñan-Pepinosa DF, Navarro-Ramirez LM, Reyes-Castellanos JE, Vivas-Giraldo DA. Debriefing-focused educational interventions in simulation-based health professions education: A systematic review of faculty and learner outcomes. Med Sci Educ. 2026;36:1669–83.
5. Tanoubi I, Labben I, Guédira S, Drolet P, Perron R, Robitaille A, Gélinas C, Bussières JS. The impact of a high fidelity simulation-based debriefing course on the Debriefing Assessment for Simulation in Healthcare (DASH) score of novice instructors. J Adv Med Educ Prof. 2019;7:159–64.
6. Cheng A, Grant V, Huffman J, Burgess G, Szyld D, Robinson T, Eppich W. Coaching the debriefer: Peer coaching to improve debriefing quality in simulation programs. Simul Healthc. 2017;12:319–25.
7. Brett-Fleegler M, Rudolph J, Eppich W, Monuteaux M, Fleegler E, Cheng A, Simon R. Debriefing Assessment for Simulation in Healthcare: Development and psychometric properties. Simul Healthc. 2012;7:288–94.
8. Muller-Botti S, Maestre JM, del Moral I, Fey M, Simon R. Linguistic validation of the Debriefing Assessment for Simulation in Healthcare in Spanish and cultural validation for 8 Spanish-speaking countries. Simul Healthc. 2021;16:13–9.
9. Olvera-Cortés HE, Hernández-Gutiérrez LS, Scherer-Castanedo E, Ramírez-Arias JD. Argueta Muñoz FD, Gutiérrez-Barreto SE: Who can Debriefing Assessment for Simulation in Healthcare? Different

perspectives on teledebriefing assessments. Simul Healthc. 2022;17:308–12.
10. 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;296:1094–102.
11. Fraundorf SH, Caddick ZA, Nokes-Malach TJ, Rottman BM. Cognitive perspectives on maintaining physicians' medical expertise: III. Strengths and weaknesses of self-assessment. Cogn Res Princ Implic. 2023;8:58.
12. Ludgate MB, Morse EE, Brown HM, Min JY, Chan AC. Gender differences in self-assessment among clerkship medical students despite equivalent academic and clinical performance. Acad Psychiatry. 2026;50:54–8.
13. Handke L, Friederichs H. Gender differences in medical students' self-assessment: Longitudinal multi-cohort study. JMIR Med Educ. 2026;12:e91071.
14. Lee-Jayaram J, Ishikawa K, Lee YJ, Tanaka LY, Lee B, Liang BX, Berg BW. Localizing and tailoring the Debriefing Assessment for Simulation in Healthcare to optimize fit. Cureus. 2024;16:e75570.
15. Al-Khayat T, Carter S, Mauger M, Patel A, Patel K, Chavarria L. Implementing the Debriefing Assessment for Simulation in Healthcare (DASH) tool for training medical faculty. Cureus. 2024;16:e69290.
16. Cooper DD, Wilson AB, Huffman GN, Humbert AJ. Medical students' perception of residents as teachers: Comparing effectiveness of residents and faculty during simulation debriefings. J Grad Med Educ. 2012;4:486–9.
17. Adams TN, Newton C, Patel H, Sulistio M, Tomlinson A, Lee W. Resident versus faculty member simulation debriefing. Clin Teach. 2018;15:462–6.
18. Klein R, Julian KA, Koch J, Snyder ED, Jassal S, Simon W, et al. Gender differences in clinical performance assessment of internal medicine residents: A longitudinal analysis of the influence of faculty and trainee gender. Acad Med. 2024;99:1413–22.
19. Gottlieb M, Chung A, Battaglioli N, Sebok-Syer SS, Kalantari A. Impostor syndrome among physicians and physicians in training: A scoping review. Med Educ. 2020;54:116–24.
20. Supplemental. Digital Content Legends.
21. Supplemental Digital Content 1. Table. Scenario-specific adjusted gender differences in the global DASH Instructor Version-Short Form score across 22 recurring clinical scenarios.
22. Supplemental Digital Content 2. Table. Scenario-specific adjusted differences in the global DASH Instructor Version-Short Form score between instructors with and without documented prior simulation training.
23. Supplemental Digital Content 3. Table. Coding framework and frequencies for self-reported prior simulation training.

Figures

Figure 1. Adjusted global and item-level differences in instructor self-assessed DASH scores by prior simulation training and gender. Negative values indicate lower ratings for instructors with any documented training versus no training (panel A) or for women versus men (panel B).



Models adjusted for academic year, harmonized clinical scenario, and the other exposure; cluster-robust 95% confidence intervals.

Figure 1

See image above for figure legend

Figure 2. Adjusted women-minus-men differences in the global DASH score within each level of prior simulation training. The interaction was exploratory.

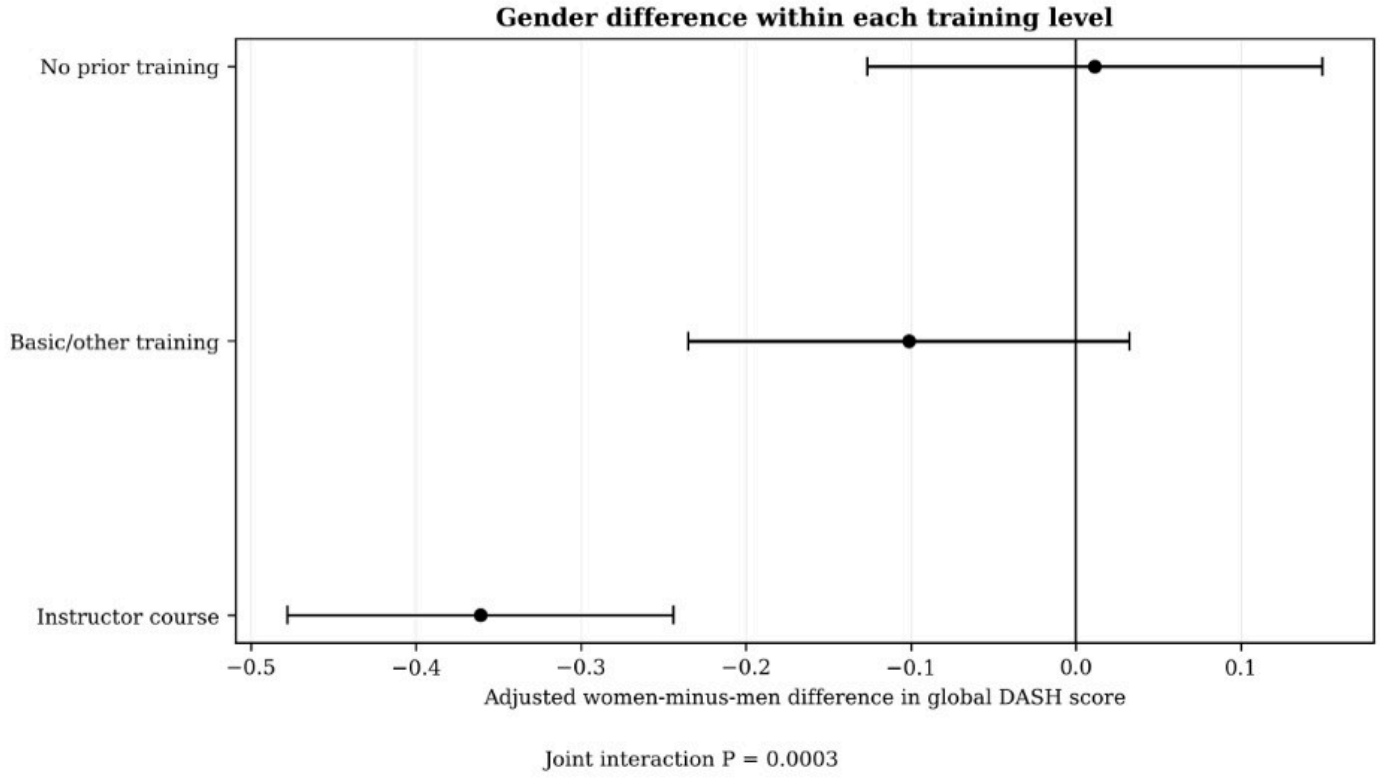
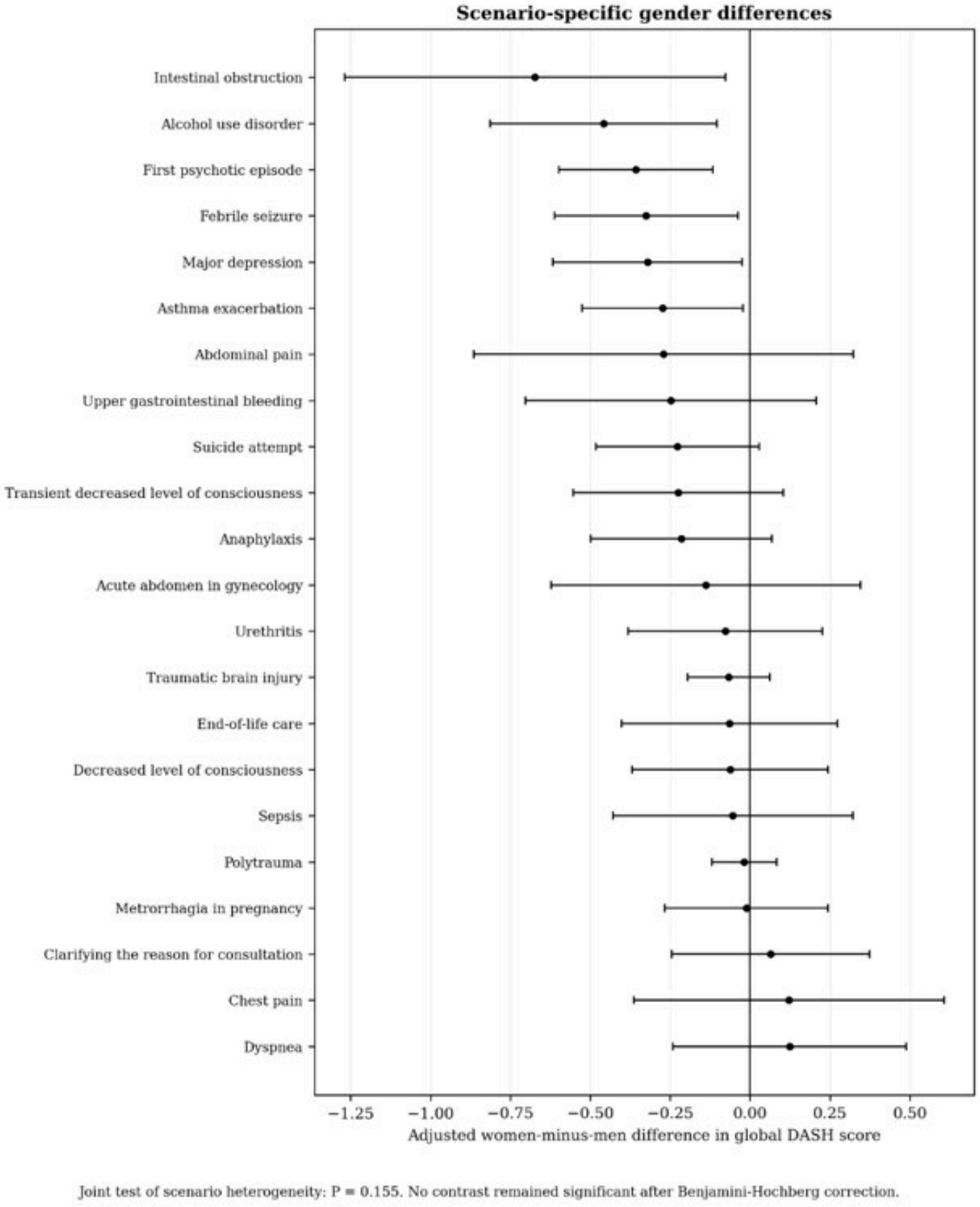


Figure 2

See image above for figure legend

Figure 3. Scenario-specific adjusted women-minus-men differences in the global DASH score across the 22 recurring clinical scenarios. No scenario-specific contrast remained significant after false-discovery-rate correction.



Joint test of scenario heterogeneity: P = 0.155. No contrast remained significant after Benjamini-Hochberg correction.

Figure 3

See image above for figure legend