2023, Volume 56, Issue 4, 234-241, e-ISSN 1938-3800

Family Medicine

ORIGINAL ARTICLE

Novel Performance Rating Instruments for Gynecological Procedures in Primary Care: A Pilot Study

Parisa Rezaiefar, MD, BSc[a]; Nisha Waqas, MD[b]; Douglas Archibald, PhD[c]; Susan Humphrey-Murto, MD, MEd[d]

AUTHOR AFFILIATIONS:

[a] Department of Family Medicine, University of Ottawa, Origyns Medical Clinic, Ottawa, ON, Canada

[b] Department of Family Medicine, University of Ottawa, Ottawa, ON, Canada

[c] Department of Family Medicine, Department of Innovation in Medical Education, Faculty of Education, Bruyère Research Institute, University of Ottawa, Ottawa, ON, Canada

[d] Department of Medicine, Department of Innovation in Medical Education, University of Ottawa, Ottawa, ON, Canada

CORRESPONDING AUTHOR:

Parisa Rezaiefar, Department of Family Medicine, University of Ottawa, Origyns Medical Clinic, Ottawa, ON, Canada, prezaief@uottawa.ca

HOW TO CITE: Rezaiefar P, Waqas N, Archibald D, Humphrey-Murto S. Novel Performance Rating Instruments for Gynecological Procedures in Primary Care: A Pilot Study. *Fam Med.* 2023;56(4):234-241.
doi: 10.22454/FamMed.2023.261011

PUBLISHED: 11 September 2023

KEYWORDS: family medicine, gynecologic procedures, procedure specific checklist

ABSTRACT

Background and Objectives: Improving training and confirming the acquisition of gynecological procedure skills for family physicians (FPs) is crucial for safe health care delivery. The objectives of this study were to (a) develop performance rating instruments for four gynecological procedures, and (b) pilot them to provide preliminary validity evidence using modern validity theory.

Methods: Sixteen academic FPs and gynecologists participated in a modified Delphi technique to develop procedure-specific checklists (PSCs) for four procedures: intrauterine device insertion, endometrial biopsy, punch biopsy of the vulva, and routine pessary care. We modified a previously validated global rating scale (GRS) for ambulatory settings. Using prerecorded videos, 19 academic FPs piloted instruments to rate one first-year and one second-year family medicine resident's performance. They were blinded to the level of training. We compared the mean scores for PSCs and GRS for each procedure using paired samples t tests and Cohen's *d* to estimate effect sizes.

Results: Consensus on items for the final PSCs was reached after two Delphi rounds. PSC and GRS scores were numerically higher for the second-year resident than the first-year resident for every procedure, with statistically significant differences for six of eight comparisons ($P<.05$). All comparisons demonstrated large effect sizes (Cohen's $d>0.8$). Both instruments received high scores for ease of use by raters.

Conclusions: We developed novel performance rating instruments for four gynecological procedures and provided preliminary validity evidence for their use for formative feedback in a simulation setting. This pilot study suggests that these instruments may facilitate the training and documentation of family medicine trainees' skills in gynecological procedures.

INTRODUCTION

Lack of access to gynecological procedures has negative consequences for patients.[1,2] Currently, specialists' availability is limited,[3,4] and family physicians (FPs) lack the skills to meet patient needs.[5,6] The College of Family Physicians of Canada (CFPC) lists 62 procedures that FPs are expected to master during their training, four of which are gynecological.[7] Among them are Pap smear, intrauterine device (IUD) insertion, endometrial aspiration biopsy, vaginal pessary fitting, and routine care. Despite this expectation, formal and standardized measures to evaluate trainees' procedural skills do not exist at the licensing level.[8] Moreover, methods such as using threshold numbers (ie, number of procedures completed in the past by a learner)[9] or self-reported confidence have been identified as poor indicators for determining individual competence.[10,11] Therefore, a global move has taken place from time-based training to competency-based medical education, where learners must demonstrate the appropriate knowledge and skills.[12,13] The Accreditation Council for Graduate Medical Education[14] and the Royal College of Physicians and Surgeons of Canada[15] follow this competency-based model.

While most procedures carry general risks such as infection and bleeding, some risks are specifically associated with gynecological procedures. For example, improper placement of an IUD can result in an unplanned pregnancy,[16] and uterine perforation requiring surgical intervention has been reported in 0.1% to 0.3% of instances of IUD insertion and endometrial biopsy.[17,18] Thus, ensuring that FPs have the necessary technical skills to perform gynecological procedures is crucial to improving safe health care delivery.

Limited training and lack of opportunities to practice have been cited as significant obstacles for family medicine

residents.[18,19] Opportunities to learn and practice procedural skills in the workplace are limited for several reasons: a short family medicine training program, work-hour limitations, and the large number of procedures on the training expectations list.[20,21] To address similar challenges in medical and surgical specialties, simulation has taken on increased importance in training[22–24] by allowing for direct observation and provision of immediate and actionable feedback in a psychologically safe environment.[25] Both low- and high-fidelity simulations have been shown to be valuable educational tools that enhance procedural skill acquisition, retention, confidence, and desire to perform a procedure.[26–28] However, evidence supporting the use of simulation to determine competence levels for real-patient procedures has remained mixed.[29,30] Regrettably, simulation remains underutilized as an educational tool in family medicine postgraduate training.[31]

The use of checklists and rating scales as part of the simulation experience facilitates procedural skills training through provision of formative and summative feedback.[32–34] Currently, no validated performance rating instruments exist for any gynecological procedures in clinical or simulated settings. When developing a new rating instrument, considering the validity of that instrument is important. The term validity refers to the degree to which the conclusions (interpretations) derived from the results of any test are "well-grounded or justifiable."[35] Validity describes how well one can legitimately trust the results of a test as interpreted for a defined intended purpose.[35] The five validity sources include content, response process, internal structure, relations to other variables, and consequence.[36] The higher the stakes of a rating instrument, the more validity evidence must be gathered.[36]

Performance rating instruments commonly used in procedural skills education include procedure-specific checklists (PSCs) and global rating scales (GRSs).[37] A PSC is comprised of a series of observable and sequential steps required to complete a technical task. Therefore, its optimal use is to provide formative feedback necessary for improving performance. A GRS targets overall psychomotor performance to evaluate trainees' ability to incorporate knowledge into task execution[38] and is a well-validated tool for summative (pass-fail) feedback.[37]

Providing feedback using performance rating instruments is a well-established educational strategy, yet family medicine lacks objective modalities to teach and evaluate the procedural skills of family medicine trainees. The first objective of this study was to develop performance rating instruments for four gynecological procedures in family medicine. Our second objective was to provide preliminary validity evidence for their use in simulation settings and for formative feedback using modern validity theory.[36]

METHODS

Procedures Selection

We chose IUD insertion, endometrial biopsy, and routine pessary care from the CFPC priority list because of their importance in providing comprehensive care.[39–44] We added punch biopsy of vulva because it is a CFPC mandatory skin procedure skill for graduates.[7,39] It can facilitate diagnosis and management of vulva conditions affecting one in five women.[45]

The Bruyère Continuing Care, Ottawa Health Science Network (OHSN), and University of Ottawa research ethics boards granted exemptions to ethics approval for this quality improvement study.

Content Development of the PSCs

Given the lack of validated PSC tools, an initial set of items were developed by two academic FPs with more than 20 years of relevant procedural and teaching experience and two academic gynecologists with medical education expertise in simulation. We used empirical evidence from a literature search of Medline, Education Source, World Wide Web, and Google Scholar, as well as the academic FPs' and gynecologists' clinical experience.[40–46] The initial PSCs had 29 items for IUD insertion, 27 for endometrial biopsy, 21 for punch biopsy of vulva, and 15 for routine pessary care.

Content Development of the GRS

We modified a previously validated GRS[24,47] for surgical skills to accommodate study procedures using partial-task models. We used the same six categories (ie, knowledge of instruments, instruments handling, respect for tissue, economy of movement, flow of procedure/forward planning, and use of assistant) and the 5-point original anchors.

We sought content evidence for the PSCs using a modified Delphi technique—a systematic means for developing consensus among a group of experts commonly used for rating instrument development.[48,49] Using the postgraduate program websites of the 17 Canadian universities with a department of family medicine, we identified 34 potential expert academic FPs and gynecologists who perform, supervise, and teach one or more of the study procedures. We sent recruitment emails using the email addresses listed on the websites.

We developed an online questionnaire for data collection, allowing participants to anonymously provide feedback on the PSCs' content. We instructed participants to rate each item based on its importance to be included in the rating instrument using an 8-point Likert scale: unimportant (1-2), somewhat unimportant (3-4), somewhat important (5-6), and very important (7-8).[50,51] We elected an even number of anchors to discourage neutral responses.[52] In addition, free-text comments were allowed for every item on the checklist to justify the participants' selections. During the first round of the Delphi, participants also were allowed to add items. To reduce cognitive load for future raters, the goal was to limit the final number of items included for each PSC to a maximum of 25.[53,54]

Data analysis was determined a priori to include items if 70% or more participants scored them 7-8 and no more than 20% scored them 1-2. We would remove items scored 1-2 by 70% of participants, with no more than 20% scoring them 7-8. All other items were to be sent for rescoring. Based on previous literature, we anticipated that two to three rounds would be sufficient.[48]

To pilot the rating instruments, we created videos of one incoming postgraduate year 1 (PGY1) and one graduating postgraduate year 2 (PGY2) family medicine resident, both females, performing each procedure in a simulation setting using partial-task models. We informed these volunteers of the procedures they would be tested on and provided them access to the instructional videos developed by the team based on the final PSCs from the modified Delphi technique for each procedure and available on YouTube (https://www.youtube.com/channel/UCBA36JDbyVhhfh7QKsocbBw). On the day of the video recording, we provided residents with written instructions for the simulation stations. The principal investigator provided passive assistance, as would a clinical supervisor, when requested by participants. The anonymity of volunteers was ensured by recording only their hands as they performed procedures. Videos ranged from 8 to 12 minutes for the PGY1 and 3 to 9 minutes for the PGY2 resident.

If academic FPs could understand and use the rating instruments with minimum training to rate the performance of the residents while watching the videos, this evidence would demonstrate support for the response process. The relation to other variables would be supported if the PGY2 resident outperformed the PGY1 resident (ie, scores would be positively associated with the level of training). In addition, given that both rating instruments measured the same construct (ie, demonstrated skills to complete the procedure successfully), we expected a positive correlation between the PSC and GRS scores.

We recruited academic FPs from one institution via the monthly departmental newsletter. We offered four workshops, with two simultaneously run sessions: one for IUD insertion and routine pessary care, and the other for endometrial biopsy and punch biopsy of vulva. Each 1-hour workshop included how to use the two rating instruments to teach the procedures and rate a resident's performance (Figure 1). Blinded to the level of training, participants watched the video of the PGY1 resident and the video of the PGY2 resident performing each procedure. They were instructed to complete the PSC first, then the GRS. In addition, participants were asked to surmise whether the resident was a PGY1 or PGY2 and to code their score accordingly.

Data Collection

The PSC and GRS scores were collected anonymously using a website tool (https://parisa-gp4pc-manuscript-demo.web.app). A nonbinary behavioral anchor of "not done" (0), "attempted/prompted" (1), and "completed" (2) was assigned to rate each item on the checklist.[42] The maximum expected score for IUD insertion was 56 (28 items), 50 (25 items) for endometrial biopsy, 40 (20 items) for punch biopsy of vulva, and 30 (15 items) for routine pessary care. The maximum expected score for the GRS was 30 (six items and a 5-point descriptive anchor) for all four procedures.

An anonymous online survey also was available on the study website to collect participants' evaluations of the rating instruments. Ease of use, visual design, comprehension, language, response anchors, and relevance to assessing residents' performance were rated using a 6-point Likert scale (strongly disagree, disagree, somewhat disagree, somewhat agree, agree, strongly agree).

Data Analysis

We compared the mean scores for the PSC and the GRS for the PGY1 and PGY2 residents for each procedure using paired sample t tests. We calculated effect sizes using Cohen's d. We also calculated Pearson correlations between PSC and GRS scores for each year of training and each procedure. We performed descriptive data analysis for the evaluation of the PSC and the GRS as rating instruments.

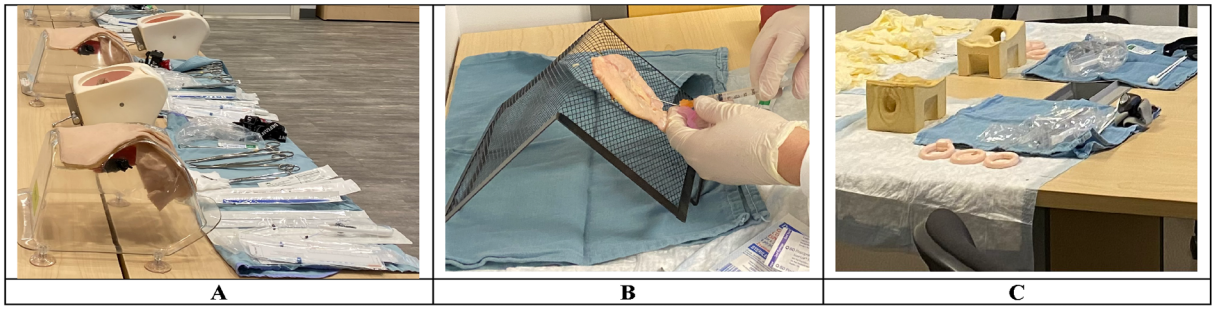
RESULTS

Twelve academic FPs and four academic gynecologists agreed to participate, representing nine Canadian universities from the provinces of Ontario, Nova Scotia, Manitoba, Alberta, and Quebec. All but one participant was female. Of the 12 FPs, four had certificates of added competence in women's health and one in dermatology. Participants provided feedback only on procedures for which they met expert criteria defined in the recruitment email. On the first round, all 16 participants completed the survey for IUD insertion, 15 completed endometrial biopsy and punch biopsy of the vulva, and 13 completed routine pessary care. Completion rates after the second round were 88% (n=14) for IUD insertion, 93% (n=14) for endometrial biopsy, 86% (n=13) for punch biopsy of the vulva, and 100% (n=13) for routine pessary care.

Table 1 summarizes the number of items on the initial list and the results of the first and second rounds of the modified Delphi technique. After the first round and based on participants' comments, items #2 (Speculum) and #3 (Lubricant) on the endometrial biopsy and IUD insertion PSCs were combined into "Lubricated or warmed speculum of appropriate size," corresponding to item #2 of the endometrial biopsy and IUD insertion PSCs. Participants also recommended to combine items "Identifies uterine position correctly when performing bimanual exam" (item #11 on the endometrial biopsy PSC and item #12 on the IUD insertion PSC) and "Checks for cervical motion tenderness when performing bimanual exam to rule out active pelvic inflammatory disease" (item #12 on the endometrial biopsy PSC and item #13 on the IUD insertion PSC) into a new item described as "Performs a bimanual exam in an attempt to determine the position of uterus (anteverted/retroverted) and reports any gross abnormality" (item #10 on the endometrial biopsy PSC and item #11 on the IUD insertion PSC). Finally, participants recommended a new item be added to the IUD insertion PSC: "Reports a normal cervix and obtains swabs for STIs if determined appropriate" (newly added item #13). Indeterminate and new items were sent on to the second round.

Consensus was reached after the second round given that we had achieved the final PSC containing 25 items for endometrial biopsy, 20 items for punch biopsy of the vulva, and 15 items for routine pessary care (https://parisa-gp4pc-manuscript-demo.web.app). For the IUD insertion, none of the

FIGURE 1. Model Design and Stations Used in the Study and the Pilot Faculty Development



A: intrauterine device insertion and endometrial biopsy; **B**: punch biopsy of vulva; **C**: routine pessary care

TABLE 1. Results of Modified Delphi Consensus for PSC Item Selection

	Intrauterine device insertion	Endometrial biopsy	Punch biopsy of vulva	Routine pessary care
Number of items to be rated	29	27	21	15
Round 1				
Items included	21	17	17	12
Items excluded	0	0	0	0
Items neither included nor excluded*	8	10	4	3
Incorporation of participants' comments	Combined 4 items into 2 modified items and added one new item	Combined 4 items into 2 modified items	Combined 2 items into 1	N/A
Items included after changes	20	16	16	12
Round 2				
Items included	3	4	2	1
Items excluded	0	0	0	0
Items neither included nor excluded after Round 2	5	5	2	2
Number of items after two rounds	28 (20+3+5)	25 (16+4+5)	20 (16+2+2)	15 (12+1+3)

An 8-point scoring anchor was used for each item: 1-2, unimportant; 3-4, somewhat unimportant; 5-6, somewhat important; 7-8, very important. Items included: >70% ranked 7-8, <20% ranked 1-2. Items excluded: >70% rated 1-2, <20% rated 7-8.
*Items sent to Round 2
Abbreviation: PSC, procedure-specific checklist

items met the a priori criteria for removal after two rounds. The closest an item came to not being important was at 14%, well below the set criteria for removal after Round 2. The team decided that stability of responses was reached[55] and that all included items were essential given the complexity of IUD insertion. To avoid false consensus, where participants agree to remove some items solely to end the process due to fatigue,[48] the team accepted that the IUD insertion PSC would remain at 28 items.

Nineteen academic FPs attended the faculty development event and rated the videos. All but two participants were female. The PSC and GRS scores were numerically higher for the PGY2 resident than for the PGY1 resident, with a large effect size (ranging from 1.03 to 4.78) for all four procedures (Table 2). Statistical significance was reached for all the PSC scores except IUD insertion and the GRS scores for punch biopsy of vulva. Table 3 demonstrates Pearson correlations between the PSC and GRS scores, none of which reached statistical significance.

Rater satisfaction was high for the PSCs and the GRS for all four procedures. As shown in Table 4, mean ratings were above 4.8 (out of 6) for ease of use, visual design, comprehension, language, response anchors, and relevance to assessing trainee's performance for both instruments.

TABLE 2. Mean Scores and Effect Sizes for PSC and GRS Scores for PGY1 and PGY2 Residents

Procedure		PGY1 resident video		PGY2 resident video			
	N	Mean	SD	Mean	SD	*P*	*d*
PSC Scores							
Intrauterine device insertion (out of 56)	6	43.50	7.89	49.50	2.26	NS	1.03
Endometrial biopsy (out of 50)	9	43.00	2.29	45.44	2.29	.002	1.09
Punch biopsy of vulva (out of 40)	5	36.80	1.64	38.40	1.14	.016	1.13
Routine pessary care (out of 30)	9	16.11	2.32	23.67	1.58	<.001	3.81
GRS Scores							
Intrauterine device insertion (out of 30)	5	15.00	3.67	24.60	2.19	.004	3.17
Endometrial biopsy (out of 30)	9	22.00	2.00	29.44	1.13	<.001	4.58
Punch biopsy of vulva (out of 30)	4	21.25	2.87	27.75	2.06	NS	2.60
Routine pessary care (out of 30)	9	14.22	4.02	28.56	1.33	<.001	4.78

Note. N=number of PSC and GRS scores completed. Cohen's d value of 0.8 or greater indicates a large effect size of paired samples.
Abbreviations: PSC, procedure-specific checklist; GRS, global rating scale; PGY, postgraduate year; NS, not statistically significant; SD, standard deviation; *d*, Cohen's *d*

TABLE 3. Pearson Correlations Between the PSC and GRS Scores for PGY1 and PGY2 Resident Videos

Procedure	Pearson correlation between PSC and GRS for PGY1	Pearson correlation between PSC and GRS for PGY2
Intrauterine device insertion	0.488	0.597
Endometrial biopsy	0.436	0.467
Punch biopsy of vulva	-0.194	-0.189
Routine pessary care	0.349	-0.79

None of the correlations between PSC and GRS scores for PGY1 and PGY2 videos were statistically significant.
Abbreviations: PSC, procedure-specific checklist; GRS, global rating scale; PGY, postgraduate year

TABLE 4. Rater Satisfaction of the PSC and GRS

Mean score (of 6) per item evaluated*							Mean score average (of 36)
	Easy to use	Visual design is acceptable	Easy to comprehend	Item is clear	Anchors are easy to use	Relevant to assessing performance	
PSC							
Intrauterine device insertion (n=6)	5.50	5.83	5.50	5.50	5.67	5.17	33.17
Endometrial biopsy (n=4)	5.60	5.67	5.33	5.50	5.33	5.17	32.66
Punch biopsy of vulva (n=4)	5.75	5.75	5.50	5.50	6.00	6.00	34.50
Routine pessary care (n=5)	5.40	6.00	4.80	4.80	5.00	5.20	31.20
GRS							
Intrauterine device insertion (n=6)	5.67	5.33	5.50	5.17	5.33	5.33	32.33
Endometrial biopsy (n=6)	5.67	5.67	5.50	5.50	5.33	5.33	32.83
Punch biopsy of vulva (n=4)	5.50	5.75	5.75	5.75	6.00	5.75	34.50
Routine pessary care (n=5)	5.00	5.80	5.00	4.80	5.00	5.40	31.00

Mean score=mean of all items listed.
*Six-point Likert scale (1=strongly disagree, 2=disagree, 3=somewhat disagree, 4=somewhat 5=agree, 6=strongly agree).
Abbreviations: PSC, procedure-specific checklist; GRS, global rating scale

DISCUSSION

The lack of specialist availability makes the education of competent FPs who provide gynecological procedures in primary care a pressing issue. Enhancing FPs' education in gynecological procedures using validated tools to teach and document technical skills is paramount to improving patient care. Using rigorous consensus methodology and modern validity theory as a framework, our team developed and piloted PSCs and a GRS for four priority gynecologic procedures for the purpose of formative feedback in a simulation setting for family medicine training.

When developing a new rating instrument in medical education, we gathered validity evidence from three of the five sources recommended by modern validity theory.[35,36] First, content evidence is supported by the fact that geographically and professionally diverse experts, based on a rigorous consensus method (modified Delphi technique), found all items relevant, confirming the comprehensiveness of the PSCs. The two-round modified Delphi technique confirmed the importance and relevance of items on PSCs. The final list for each procedure included 15 to 28 items, a manageable number for raters. Though the final list for IUD insertion was not reduced to 25 items as was set a priori, the faculty development pilot participants did not negatively evaluate this PSC. Second, the raters' high evaluation of the two rating instruments for all four procedures illustrates evidence for the response process. The minimal rater training required in the pilot suggests ease of future implementation. Lastly, the relation to other variables was supported by the higher scores acquired for all four procedures by the PGY2 resident compared to the PGY1 resident, with impressive effect sizes.

The small sample size can explain the observed high standard deviation from the mean. Positive correlations between the PSC and GRS scores could not be demonstrated due to a lack of statistical significance. The primarily positive trends would be considered moderate, although negative trends were noted between the PSC and GRS scores for the punch biopsy of vulva performed by both residents and for the routine pessary care performed by the PGY2. Therefore, results must be interpreted with caution. Nevertheless, our findings are consistent with a systematic review of 16 studies documenting considerable inconsistency in the correlation between PSC and GRS scores, with a moderate trend reported at best.[38]

This study offers two potential educational opportunities for family medicine educators. First, the PSCs developed in this study required minimum training for educators to use, and the models used were reusable, portable, and inexpensive. Therefore, educators may use these simulation-based tools for trainees to practice and receive task-specific corrective feedback, reinforce correct actions, and increase their motivation. Such teaching strategies can offer an excellent solution for improving patient experiences and compensating for short family medicine training schemes and ultimately may enhance trainees' proficiency.[56] Future studies to examine the correlation between the scores in simulation and clinical setting and patient outcomes are warranted.

Second, trainees report significant anxiety about performing gynecological exams.[57,58] The stress experienced by trainees during simulation-based education improves memory and skills retention.[59] The PSCs developed in this study, combined with simulation, can offer an excellent tool for trainees to improve their technical skills by experiencing the stress associated with such procedures outside of the clinical setting.[60] Anxiety reduction and improved confidence after simulation are well-documented by the literature and are associated with increased willingness to perform a procedure.[60,61]

Because our instruments were piloted at a single institution using the videos of only one PGY1 and one PGY2 resident, our ability to test inter- and intrarater reliability was limited; therefore, our findings may not be generalizable to other settings. Further research to improve the PSC anchors by splitting the prompted/attempted anchors to individual ones can be considered. Evidence for internal structure and consequence validity must be gathered before the rating instruments developed in this study can be used for the purpose of summative feedback and as assessment tools. Building on this study, we plan to test these instruments in a larger sample size of trainees and raters, and to explore whether learning these procedural skills in a simulated setting can predict performance in the workplace.

CONCLUSIONS

We developed two performance rating instruments for four essential gynecological procedures in family medicine and provided preliminary validity evidence for their use for formative feedback in a simulation setting. The rating instruments and simulations designed in this pilot study can be easily incorporated into the postgraduate curricula for teaching and documenting the acquisition of gynecological skills in family medicine. Further validation is required if these tools are to be used for more summative assessment purposes.

ACKNOWLEDGMENTS

We acknowledge the following individuals for their contribution to this project. Drs Christiane Kuntz, Glenn Posner, and Adam Garber helped in the development of initial PSCs content. Drs William Carron, Emily Breecher, Taunia Rifai, Christiane Kuntz, Stefanie Vescio, Emily Devitt, Sakshi Mahajan Malik, and Elise Azzi were instructors at the pilot faculty development session. Mr Firas Jribi, Ms Vanshika Chaudhary, and Dr Sakshi Mahajan Malik developed the website platform for the data collection.

REFERENCES

1. Socías ME, Koehoorn M, Shoveller J. Gender inequalities in access to health care among adults living in British Columbia. *Canada. Womens Health Issues.* 2016;26(1):74-79.
2. Pulice-Farrow L, Lindley L, Gonzalez KA. Wait, what is that? A man or woman or what?": trans microaggressions from gynecological healthcare providers. *Sex Res Soc Policy.* 2022;19(4):549-550.

3. Liddy C, Nawar N, Moroz I. Understanding patient referral wait times for specialty care in Ontario: A retrospective chart audit. *Healthc Policy.* 2018;13(3):59-69.
4. Fryer GE, Green LA, Dovey SM, Phillips RI. The United States relies on family physicians unlike any other specialty. *Am Fam Physician.* 2001;63(9):1669.
5. Coffman M, Wilkinson E, Jabbarpour Y. Despite adequate training, only half of family physicians provide women's health care services. *J Am Board Fam Med.* 2020;33(2):186-188.
6. Chelvakumar M, Shaw JG. Trained and ready, but not serving? Family physicians' role in reproductive health care. *J Am Board Fam Med.* 2020;33(2):182-185.
7. Fowler N, Wyman R. *Residency Training Profile for Family Medicine and Enhanced Skills Programs Leading to Certificates of Added Competence.* College of Family Physicians of Canada; 2021. https://www.cfpc.ca/CFPC/media/Resources/Education/Residency-Training-Profile-ENG.pdf.
8. Crichton T, Schultz K, Lawrence K. *Assessment Objectives for Certification in Family Medicine.* 2020. https://www.cfpc.ca/en/education-professional-development/educational-frameworks-and-reference-guides/assessment-objectives-for-certification-in-fm.
9. Ekkelenkamp VE, Koch AD, Man RAD, Kuipers EJ. Training and competence assessment in GI endoscopy: a systematic review. *Gut.* 2016;65(4):607-615.
10. Davis DA, Mazmanian PE, Fordis M, Harrison RV, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA.* 2006;296(9):94-95.
11. Rezaiefar P, Forse K, Burns JK. Does general experience affect self-assessment?. *Clin Teach.* 2019;16(3):197-202.
12. Kirkpatrick D, Kirkpatrick J. *Evaluating Training Programs: The Four Levels*. Koehler Publishers; 2006. .
13. Frank JR, Snell LS, Cate OT. Competency-based medical education: theory to practice. *Med Teach.* 2010;32(8):638-645.
14. Family Medicine Milestones. *Accreditation Council for Graduate Medical Education.* 2015. https://www.acgme.org/globalassets/pdfs/milestones/familymedicinepreventivemedicinemilestones.pdf.
15. Competence by Design: Canada's Model for Competency-based Medical Education. *Royal College of Physicians and Surgeons of Canada.* 2023. https://www.royalcollege.ca/ca/en/cbd.html.
16. Johnson BA. Insertion and removal of intrauterine devices. *Am Fam Physician.* 2005;71(1):95-102.
17. Markovitch O, Klein Z, Gidoni Y, Holzinger M, Beyth Y. Extrauterine mislocated IUD: is surgical removal mandatory?. *Contraception.* 2002;66(2):105-108.
18. Williams PM, Gaddey HL. Endometrial biopsy: tips and pitfalls. *Am Fam Physician.* 2020;101(9):551-556.
19. Mackenzie MS, Berkowitz J. Do procedural skills workshops during family practice residency work?. *Can Fam Physician.* 2010;56(8):296-301.
20. Tiemstra JD. Fixing family medicine residency training. *Fam Med.* 2004;36(9):666-668.
21. Sigmon JL, Mcpherson V, Little JM. Four years of training in family medicine: implications for residency redesign. *Fam Med.* 2012;44(8):550-554.
22. Rosen KR. The history of medical simulation. *J Crit Care.* 2008;23(2):157-166.
23. Barsuk JH, Ahya SN, Cohen ER, Mcgaghie WC, Wayne DB. Mastery learning of temporary hemodialysis catheter insertion by nephrology fellows using simulation technology and deliberate practice. *Am J Kidney Dis.* 2009;54(1):70-76.
24. Reznick R, Regehr G, Macrae H, Martin J, Mcculloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg.* 1997;173(3):226-230.
25. Ennen CS, Satin AJ. Training and assessment in obstetrics: the role of simulation. *Best Pract Res Clin Obstet Gynaecol.* 2010;24(6):747-758.
26. Sawyer T, White M, Zaveri P. Learn, see, practice, prove, do, maintain: an evidence-based pedagogical framework for procedural skill training in medicine. *Acad Med.* 2015;90(8):33-33.
27. Dehmer JJ, Amos KD, Farrell TM, Meyer AA, Newton WP, Meyers MO. Competence and confidence with basic procedural skills: the experience and opinions of fourth-year medical students at a single institution. *Acad Med.* 2013;88(5):682-687.
28. Stewart RA, Hauge LS, Stewart RD, Rosen RL, Charnot-Katsikas A, Prinz RA. Association for Surgical Education. A CRASH course in procedural skills improves medical students' self-assessment of proficiency, confidence, and anxiety. *Am J Surg.* 2007;193(6):771-773.
29. Sarker SK, Albrani T, Zaman A, Kumar I. Procedural performance in gastrointestinal endoscopy: live and simulated. *World J Surg.* 2010;34(8):764-765.
30. Park J, Macrae H, Musselman LJ. Randomized controlled trial of virtual reality simulator training: transfer to live patients. *Am J Surg.* 2007;194(2):205-211.
31. Tenegra JC, Hoffman MR, Luckey GSM, Dilalla LF, Ledford C. Simulation-based medical education in family medicine residencies: A CERA study. *Fam Med.* 2022;54(4):264-269.
32. Pugh D, Touchie C, Wood TJ, Humphrey-Murto S. Progress testing: is there a role for the OSCE?. *Med Educ.* 2014;48(6):623-631.
33. Pugh D, Halman S, Desjardins I, Humphrey-Murto S, Wood TJ. Done or almost done? Improving OSCE checklists to better capture performance in progress tests. *Teach Learn Med.* 2016;28(4):406-414.
34. Martin JA, Regehr G, Reznick R. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg.* 1997;84(2):273-278.
35. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med.* 2006;119(2).
36. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ.* 2003;37(9):830-837.
37. Cunnington JP, Neville AJ, Norman GR. The risks of thoroughness: reliability and validity of global ratings and checklists in an OSCE. *Adv Health Sci Educ Theory Pract.* 1996;1(3):227-233.
38. Ilgen JS, Ma IW, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med Educ.* 2015;49(2):161-173.
39. Nothnagle M, Sicilia JM, Forman S. Required procedural training in family medicine residency: a consensus statement. *STFM Group on Hospital Medicine and Procedural Training.*

2008;40:248-252.
40. Black A, Guilbert E, Costescu D. Canadian contraception consensus (part 1 of 4). *J Obstet Gynaecol Can.* 2015;37(10):936-942.
41. Renaud MC, Le TSGS, Policy P, Committee G. Epidemiology and investigations for suspected endometrial cancer. *J Obstet Gynaecol Can.* 2013;35(4):380-381.
42. Harvey MA, Lemieux MC, Robert M, Schulz JA. Guideline no. 411: vaginal pessary use. *J Obstet Gynaecol Can.* 2021;43(2):255-266.
43. Swift SE. The distribution of pelvic organ support in a population of female subjects seen for routine gynecologic health care. *Am J Obstet Gynecol.* 2000;183(2):277-285.
44. Busing N, Newbery P. Robust description of family practice. A look at the National Physician Survey. *Can Fam Physician.* 2005;51(5):647-649.
45. Eva LJ. Screening and follow up of vulval skin disorders. *Best Pract Res Clin Obstet Gynaecol.* 2012;26(2):175-188.
46. Sulik SM, Heath CB. *Primary Care Procedures in Women's Health*. Springer; 2010. .
47. Malacarne DR, Escobar CM, Lam CJ, Ferrante KL, Szyld D, Lerner VT. Teaching vaginal hysterectomy via simulation: creation and validation of the Objective Skills Assessment Tool for simulated vaginal hysterectomy on a task trainer and performance among different levels of trainees. *Female Pelvic Med Reconstr Surg.* 2019;25(4):298-304.
48. Humphrey-Murto S, Varpio L, Wood TJ. The use of the Delphi and other consensus group methods in medical education research: A review. *Acad Med.* 2017;92(10):491-492.
49. Custer R, Scarcella J, Stewart B. The modified Delphi technique-a rotational modification. *J Vocat Tech Educ.* 1999;15(2):50-58.
50. Oaster T. Number of alternatives per choice point and stability of Likert-type scales. *Percept Mot Skills.* 1989;68(2).
51. Preston CC, Colman AM. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychol (Amst).* 2000;104(1):1-15.
52. Chyung SY, Roberts K, Swanson I, Hankinson A. Evidence-based survey design: the use of a midpoint on the Likert scale. *Perform Improv.* 2017;56:15-23. .
53. Tavares W, Eva KW. Exploring the impact of mental workload on rater-based assessments. *Adv Health Sci Educ Theory Pract.* 2013;18(2):291-303.
54. Tavares W, Ginsburg S, Eva KW. Selecting and simplifying: rater performance and behavior when considering multiple competencies. *Teach Learn Med.* 2016;28(1):41-51.
55. Diamond IR, Grant RC, Feldman BM. Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. *J Clin Epidemiol.* 2014;67(4):401-409.
56. Aydın A, Ahmed K, Abe T. Effect of simulation-based training on surgical proficiency and patient outcomes: a randomised controlled clinical and educational trial. *Eur Urol.* 2022;81(4):385-393.
57. Rees CE, Wearn AM, Vnuk AK, Bradley PA. Don't want to show fellow students my naughty bits: medical students' anxieties about peer examination of intimate body regions at six schools across UK, Australasia and Far-East Asia. *Med Teach.* 2009;31(10):921-927.
58. Mudd JW, Siegel RJ. Sexuality-the experience and anxieties of medical students. *N Engl J Med.* 1969;281(25):397-398.
59. Demaria S, Levine AI, Levine AI, DeMaria S, Schwartz AD SimAJ. *The use of stress to enrich the simulated environment*. Springer; 2013. .
60. Yu JH, Chang HJ, Kim SS. Effects of high-fidelity simulation education on medical students' anxiety and confidence. *PLoS One.* 2021;16(5):251078.
61. Hays RB, Jolly BC, Caldon LJ. Is insight important? measuring capacity to change performance. *Med Educ.* 2002;36(10):965-971.