PRiMER
peer-reviewed reports in medical education research
research brief

Evaluating Resident Procedural Skills: Faculty Assess a Scoring Tool

Jack Wells, MD, MHA | Alicia Ludden-Schlatter, MD, MSAM | Robin L. Kruse, PhD, MSPH | Nikole J. Cronk, PhD

PRiMER. 2020;4:4.

Published: 4/17/2020 | DOI: 10.22454/PRiMER.2020.462869

Abstract

Background and Objective: Procedural skills assessment is critical in residency training. The Council of Academic Family Medicine recommends the Procedural Competency Assessment Tool (PCAT) for assessing procedure competence of family medicine residents. We sought to evaluate the reliability of the PCAT and to better identify its strengths and limitations.

Methods: In this mixed-methods study conducted in 2017, 18 faculty members of an academic family medicine residency program watched a video of one of the authors performing a simulated shave biopsy with intentional errors. Faculty scored the procedure using the shave biopsy PCAT, then participated in a focus group discussion of the rationale for the scores given. Qualitative analysis assessed perceived benefits and challenges of the PCAT. Following the discussion, faculty scored the same procedure again, using a PCAT modified with additional objective criteria.

Results: On the original PCAT, 40% of respondents rated the physician as competent. This dropped to 21.4% on the modified PCAT (*P*=.035). Respondents scored competent even though procedure components were scored as novice. Score variability decreased with the checklist-based PCAT. Qualitative analysis revealed that the PCAT is subjective and interpretation of the tool varies widely.

Conclusions: Further studies regarding PCAT validity and reliability are needed. The PCAT may require further norming with additional objective criteria to improve reliability. Residencies may train faculty on using the PCAT to improve interobserver agreement, or decide to use a more intuitive checklist evaluation tool.

Introduction

Procedural skills assessment is an important part of residency training. Although residency programs use different assessment tools,[1-3] the Council of Academic Family Medicine (CAFM) recommends the Procedural Competency Assessment Tool (PCAT).[4] The PCAT rates skill levels as novice, competent, or expert using a 5-point scale. It was adapted from the Operative Performance Rating System (OPRS), a validated global rating scale (GRS) evaluation tool for surgical procedures.[5-7] GRS assess overall performance, are influenced by observers' overall impressions, and provide a subjective interpretation of skill level.[8] Conventionally, GRS have been thought to have greater reliability than checklists, though newer literature challenges this perception.[9, 10] GRS require more evaluator training, whereas checklists are more intuitive and may have higher interrater reliability.[10] The OPRS acknowledges the likelihood of interrater variation and recommends "at least 10 different expert raters to rate each resident in each year [to] control for these rating idiosyncrasies."[6] This can be time intensive, faculty intensive, and impractical. We

recognized the need for reliable procedural skill assessment to more accurately and uniformly assess residents' skills.

We conducted a mixed-methods study to assess the reliability of the PCAT and obtain faculty feedback. We compared reliability of the PCAT to a PCAT modified with objective criteria. A qualitative assessment of a faculty focus group elucidated PCAT strengths and limitations.

Methods

During a faculty seminar in 2017, 18 family medicine faculty watched a video of one of the authors (A.L.) performing a simulated shave biopsy with intentional errors. Faculty scored the procedure using four components of the PCAT for shave biopsy,[5] condensed for time and relevance to a simulation setting: informed consent, procedure setup (learner used the wrong needle to draw lidocaine and contaminated the injecting needle); local anesthesia (learner swiveled the injecting needle underneath the skin, bending the needle); and procedure flow and efficiency (learner was inefficient, mishandled the blade by using two hands and not stabilizing the biopsy site). Figure 1 shows the PCAT scoring tool. Each component could be rated as novice (we numerically scored as 1 point), novice-plus (1.5 points), competent (2 points), competent-plus (2.5 points), or expert (3 points). Competent scores in every element yielded a minimum total score of 8; the maximum possible score was 12. The PCAT also includes a yes/no assessment of overall competence. The PCAT does not provide evaluation instructions.[4] After scoring, faculty participated in a focus group discussion regarding the rationales for their scores. Following the discussion, faculty rescored the procedure using a PCAT modified with objective criteria (Table 1).

We entered scores into a spreadsheet and imported into SAS for Windows 9.4 (SAS Institute, Inc, Cary, NC). We calculated descriptive statistics (simple frequencies, means, medians, and 95% confidence intervals [CI]). We compared item scores and the total score before and after the discussion and use of the modified PCAT using the Wilcoxon Rank Sum Test, a nonparametric test to determine whether two related samples are drawn from the same distribution. We compared participants' ratings of overall competence before and after the intervention with Fisher exact test.

Author N.C. analyzed transcripts of the focus group to identify common themes. Authors A.L. and J.W. reviewed this analysis and disagreements were resolved by consensus. The University of Missouri Institutional Review Board exempted this study.

Results

In general, scores were higher on the standard PCAT than the modified PCAT (Table 2). The scores for informed consent and administering local anesthesia were significantly lower on the modified PCAT ($P<.01$). On the standard PCAT, 40% of respondents deemed the resident competent overall. This dropped to 21.4% with the modified PCAT ($P=.035$).

Mean and median scores were lower on the modified PCAT for all items except median procedure setup. Informed consent and administering local anesthesia had statistically significant differences ($P<.01$). Mean total score dropped from 7.5 to 6.0, and the median score dropped from 7.0 to 6.0 ($P=.019$). Scores for the modified PCAT were less variable; the 95% CI was narrower for all modified PCAT component and total scores than for the standard PCAT (Table 2).

Using the standard PCAT, six of 18 faculty (33.3%) deemed the performance as competent overall, nine (50%) responded not competent, and three did not respond. Of the six who scored the learner as competent overall, three had total scores under 8, the minimum score for competence in all components. One faculty member scored all areas in the competent range, but responded "no" to the overall competence question. On the modified PCAT, three (16.7%) evaluated the performance as competent, 11 (61.1%) evaluated as not competent, and three did not answer. Of those who said competency was achieved, two of the three did not assign a minimum passing total score. One

evaluator did not complete the modified PCAT.

Following initial scoring with the standard PCAT, participants were asked to explain the rationales for their scores. Three themes emerged: confusion regarding how to use the PCAT tool, varying assumptions about the terminology (eg, novice vs competent), and how to evaluate observed behaviors. Regarding the format of the PCAT, several faculty assumed the layout implied a 3-point scale (novice, competent, or expert only) and expressed they would have scored differently if they had known to use a 5-point scale (Figure 1). One participant was unsure whether to consider the needle contamination as part of procedure setup or anesthesia.

Participants voiced different definitions of competence. One participant scored a component with an error as competent because it was "not necessarily perfect, but good enough," whereas another expressed that any error would result in a novice score for that component. Some faculty rated competence based on the end result (competent: the specimen was obtained) and some considered the means to that end (novice: the blade was held incorrectly). Some faculty adjusted their definition of competence according to the learner's level of training (eg, "expected for an intern").

Some participants used similar evidence to justify different ratings. For example, when assessing informed consent, one faculty participant gave a novice score because they didn't hear an explanation of the possible need for additional surgery, whereas another participant scored it as competent because they assumed that conversation had happened previously.

Several faculty expressed confusion regarding interpretation of the PCAT descriptors, stating that the PCAT is "subjective" or "generic." Overall, faculty stated that the PCAT is likely not reliable and that more objective anchors would be useful in interpreting competence and evaluating residents according to a consistent standard (Table 3).

Discussion

Our modified PCAT with objective criteria reduced scoring variability and also reduced inappropriate designation of "competent." Faculty using the current PCAT inappropriately evaluated procedure performance favorably and were more likely to designate a novice performance as competent. They were also more likely to evaluate a learner as competent despite scoring individual procedure components as not competent. Both of these issues improved and interrater variability decreased with the addition of objective criteria. During discussion, faculty identified challenges with the PCAT such as vague descriptors and unclear definitions of competence, and requested objective anchors for guidance.

This study is limited by small sample size at a single residency program, and may not be generalizable. The faculty watched the procedure video once, therefore the scores for the modified PCAT could have been affected by recall. Faculty were not given the modified scoring tool until after the focus group to avoid biasing the discussion, but the discussion could have influenced subsequent scoring on the modified PCAT. This indicates that group training sessions may improve reliability. Faculty requested more specific criteria to aid assessment of competency. This confusion indicates that the PCAT may benefit from additional norming (the process in which educators assess and calibrate a rubric via a discussion leading to an "evidence-driven consensus"[11]). The PCAT may require more objective criteria, clearer definitions of competency, and more explicit instructions for use.

We assessed PCAT performance for one procedure, and it may perform better for others. However, CAFM's description of the PCAT development process does not indicate that it has undergone any testing for validity or reliability[4]; formal validity and reliability studies are recommended. The PCAT may benefit from additional norming by content experts. It is possible that for evaluating procedure performance, checklists may be a better tool than GRS like the PCAT. Checklists mark defined components as either done or not done, with less evaluator subjectivity, and can be more intuitive to use.[9, 11-13.]

Conclusion

Our study revealed poor accuracy and validity of the PCAT, which is currently the recommended tool to evaluate family medicine resident procedure competence. Adding objective criteria to the PCAT improved accuracy and reliability. Qualitative assessment identified challenges in interpreting vague descriptors. This study identifies the need for formal validity and reliability assessment of the PCAT, preferably with larger sample sizes and control comparators. The PCAT may be improved by additional norming. CAFM and individual residencies may consider PCAT training sessions, or forgoing GRS tools like the PCAT in favor of checklist-based assessments.

Tables and Figures

Table 1: Anchors Used to Modify the PCAT

PCAT Area of Evaluation	Anchors Added
Informed consent	• Lists at least two potential complications (pain, infection, bleeding, scarring, transected melanoma) • Lists at least two indications for procedure (cosmetic removal, diagnosis) • Offers at least one alternative to procedure (referral, observation)
Procedure setup	• Resident familiar with (and selects) equipment appropriately (Dermablade/flexible scalpel, nonsterile gloves, alcohol, silver nitrate/aluminum chloride, bandage) • Cleans operative site with alcohol swab
Local anesthesia	• Verbalizes what anesthetic they are using (lidocaine with or without epi) • Verbalizes what size needle (22 to 25 gauge) • Injects lidocaine with bevel up, making superficial wheal
Procedure flow and efficiency	• Appropriately stabilize skin lesion to be biopsied, handle Dermablade, obtain specimen • Appropriately place specimen in preservative • Appropriately manage bleeding via gauze and silver nitrate/aluminum chloride • Apply bandage • Verbalize wound care instructions (wash with mild soap and water four times per day, Vaseline as needed for moisture, bandage as needed to contain Vaseline)

Abbreviation: PCAT, Procedural Competency Assessment Tool

Table 2: Comparison Competence Ratings in Individual Procedure Components and Overall Procedure Performance, Before and After Intervention [a]

	Before Intervention [a]		After Intervention [a]		
Component and Expertise Level	**N [b]**	**% [c]**	**N**	**%**	***P* Value**
Obtaining informed consent					.0014[d]
Novice or novice +			5	29.4	
Competent or competent +	9	50.0	12	70.6	
Expert	9	50.0			
Procedure setup					.79[d]
Novice or novice +	11	61.1	13	76.5	
Competent or competent +	5	27.8	4	23.5	
Expert	2	11.1			
Local anesthesia					.0092[d]
Novice or novice +	5	27.8	16	94.1	
Competent or competent +	12	66.7	1	5.9	
Expert	1	5.6			
Procedure flow and efficiency					.15[d]
Novice or novice +	8	44.4	15	88.2	
Competent or competent +	9	50.0	2	11.8	
Expert	1	5.6			
Overall competence					.035[e]
No	9	60.0	11	78.6	
Yes	6	40.0	3	21.4	

Mean and Median Ratings of Procedural Competency Components and Overall					
	Before Intervention		After Intervention		
Component Rated	**Mean (95% CI)**	**Median**	**Mean (95% CI)**	**Median**	***P* Value [d]**
Obtaining informed consent	2.50 (2.24-2.76)	2.5	1.85 (1.68-2.03)	2.0	.0014
Procedure setup	1.56 (1.19-1.92)	1.0	1.47 (1.28-1.66)	1.5	.79
Local anesthesia	1.81 (1.52-2.09)	2.0	1.32 (1.17-1.48)	1.5	.0092
Procedure flow and efficiency	1.64 (1.35-1.93)	2.0	1.35 (1.13-1.57)	1.5	.15
Total score	7.50 (6.57-8.43)	7.0	6.00 (5.50-6.50)	6.0	.019

[a] "Intervention" refers to focus group discussion and re-scoring procedure using modified checklist PCAT.
[b] Three participants did not rate overall competence before the intervention. After the intervention, one participant left all items blank and four participants did not rate overall competence after the intervention.
[c] Percentages within each category may not add up to 100% due to rounding.
[d] Wilcoxon rank sum test comparing scores before and after the discussion
[e] Fisher exact test

Figure 1: PCAT Scoring Tool Used in the Initial Assessment

Instructions for evaluator:

Please circle the descriptor corresponding to the candidate's performance in each category, *irrespective of the training level.*

Indication/Informed Consent:

Novice		Competent		Expert
Not sure of the patient's history, context of the procedure, or has knowledge gaps in procedure contraindications or potential complications		Understands the general indications, contraindications, potential complications, and clinical value of procedure; able to explain to patient		Clearly articulates the clinical value, potential complications, and alternatives to patient; able to accurately answer all patient questions to obtain informed consent

Procedure Setup:

Novice		Competent		Expert
Does not gather required supplies, poor patient positioning, poor sterile technique, or does not properly identify landmarks		Gathers key instruments and supplies; properly positions patient; identifies landmarks; maintains sterile technique		Anticipates supplies needed for unexpected complications; ergonomic setup of all instruments and supplies

Anesthesia:

Novice		Competent		Expert
Requires guidance to provide adequate block		Performs field block without guidance and with good anesthesia		Smoothly and efficiently performs field block without guidance and with good anesthesia

Procedure Flow and Efficiency:

Novice		Competent		Expert
Frequently stops procedure and seems unsure of next move, or many unnecessary moves		Demonstrates some forward planning with reasonable progression of procedure; efficient time/motion but some unnecessary moves		Obviously plans course of procedure with effortless flow from one move to the next; clear economy of movement and maximum efficiency

Overall, on this task did the provider demonstrate competency to perform this procedure independently?

Yes _____ No _____

Comments:

Table 3: Themes and Comments From Focus Group Discussion of Scoring Rationale Based on the Unmodified PCAT

Theme	Representative Comment
Unclear factors for determining a score of "competent"	"I…saw competent as we can work on these (issues); I still want to be [present for the procedure]. Expert would be: I would be comfortable with them on their own without me." "This is what would be expected from an intern. If I saw a third-year performing this procedure I'd be a lot more worried."
Clarity of the scoring tool	"Don't think it's valid or reliable - would not be reliable. Too subjective." "Descriptors—novice, competent, expert—are generic. Not specific to the procedures. There is not a list of things you need to observe." "When we have a field with anchors I am going to use your anchors and be as objective as I can about that."

Acknowledgments

The authors acknowledge Gwendolyn Wilson for her invaluable help in the preparation of this manuscript.

Corresponding Author

Jack Wells, MD, MHA

M245 Medical Sciences Building, Department of Family and Community Medicine, School of Medicine, University of Missouri, Columbia, MO 65212. 573-999-6502. Fax: 573-642-3015.

wellsjack@health.missouri.edu

Author Affiliations

Jack Wells, MD, MHA - Department of Family and Community Medicine, School of Medicine, University of Missouri, Columbia, MO

Alicia Ludden-Schlatter, MD, MSAM - Department of Family and Community Medicine, School of Medicine, University of Missouri, Columbia, MO

Robin L. Kruse, PhD, MSPH - Department of Family and Community Medicine, School of Medicine, University of Missouri, Columbia, MO

Nikole J. Cronk, PhD - Department of Family and Community Medicine, University of Missouri School of Medicine, Columbia, MO

References

1. Kedian T, Gussak L, Savageau JA, et al. An ounce of prevention: how are we managing the early assessment of residents' clinical skills?: A CERA study. Fam Med. 2012;44(10):723-726.
2. Whitehead C, Kuper A, Hodges B, Ellaway R. Conceptual and practical challenges in the assessment of physician competencies. Medical Teacher. 2015; 37:3, 245-251.
3. Sawyer T, White M, Zaveri P, et al. Learn, see, practice, prove, do, maintain: an evidence-based pedagogical framework for procedural skill training in medicine. Acad Med. 2015;90(8):1025-1033. https://doi.org/10.1097/ACM.0000000000000734
4. Association of Family Medicine Residency Directors. CAFM Consensus Statement for Procedural Training in Family Medicine Residency. Leawood, KS: Association of Family Medicine Residency Directors (AFMRD). 2017. https://afmrd.socious.com/page/procedures. Accessed March 29, 2017.
5. Larson JL, Williams RG, Ketchum J, Boehler ML, Dunnington GL. Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. Surgery. 2005;138(4):640-647. https://doi.org/10.1016/j.surg.2005.07.017
6. American Board of Surgery. A User's Manual for the Operative Performance Rating System (OPRS). 2012.

http://www.absurgery.org/xfer/assessment/oprs_user_manual.pdf. Accessed August 16, 2015.
7. Benson A, Markwell S, Kohler TS, Tarter TH. An operative performance rating system for urology residents. J Urol. 2012;188(5):1877-1882. https://doi.org/10.1016/j.juro.2012.07.047
8. Kim J, Neilipovitz D, Cardinal P, Chiu M. A comparison of global rating scale and checklist scores in the validation of an evaluation tool to assess performance in the resuscitation of critically ill patients during simulated emergencies (abbreviated as "CRM simulator study IB"). Simul Healthc. 2009;4(1):6-16. https://doi.org/10.1097/SIH.0b013e3181880472
9. 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. https://doi.org/10.1111/medu.12621
10. Wood TJ, Pugh D. Are rating scales really better than checklists for measuring increasing levels of expertise? Med Teach. 2020;42(1):46-51. https://doi.org/10.1080/0142159X.2019.1652260
11. Schoepp K, Danaher M, Krnov AA. An effective rubric norming process. Practical Assessment, Research, and Evaluation. 2018;(23):Article 11.
12. Turner K, Bell M, Bays L, et al. Correlation between global rating scale and specific checklist scores for professional behavior of physical therapy students in practical examinations. Educ Res Int. 2014;2014:1-6. https://doi.org/10.1155/2014/219512
13. Kogan J, Hess B, Conforti L, et al. What drives ratings of residents' own clinical skills? The impact of faculty's own skills? Acad Med. 2010;85:S25-S28. https://doi.org/10.1097/ACM.0b013e3181ed1aa3