2026, Volume 58, Issue 8, 582-585, e-ISSN 1938-3800

Family Medicine

BRIEF REPORT

Limited Agreement Regarding Outpatient Preceptor Behaviors: Observer, Residents, Self

Brandon H. Hidaka, MD, PhD[a]; Terri Nordin, MD[a]; Hannah J. Kolarik, BS[b]; Deirdre Paulson, PhD[a]

AUTHOR AFFILIATIONS:

[a]Department of Family Medicine, Mayo Clinic Health System, Eau Claire, WI

[b]Medical College of Wisconsin–Central Wisconsin, Wausau, WI

CORRESPONDING AUTHOR:

Deirdre Paulson, Department of Family Medicine, Mayo Clinic Health System, Eau Claire, WI, Paulson-ODonovan.Deirdre@mayo.edu

HOW TO CITE: Hidaka BH, Nordin T, Kolarik HJ, et al. Limited Agreement Regarding Outpatient Preceptor Behaviors: Observer, Residents, Self. *Fam Med*. 2026;58(8):582-585. doi: 10.22454/FamMed.2026.238874

FIRST PUBLISHED: September 4, 2026

KEYWORDS: clinical teaching, faculty development

ABSTRACT

Background and Objectives: We developed the Mayo Outpatient Precepting Evaluation Tool (MOPET) to assess and improve clinical teaching for family medicine residents. Given the scheduling challenges of peer evaluation, we sought to improve the MOPET's usefulness by evaluating the reliability of self- and resident-reported teaching behaviors during clinical precepting compared to peer observation.

Methods: Forty-eight half-day sessions in a family medicine residency clinic were assessed with parallel observer, preceptor, and resident MOPET forms. Interrater reliability was calculated using the κ (kappa) statistic.

Results: Residents reported that preceptors used 14 of 20 teaching behaviors more often than observers. Observers and preceptors significantly agreed on the use of five teaching behaviors, while residents agreed with observers on two teaching behaviors.

Conclusions: Although self-assessment and resident-based feedback of preceptor teaching behavior may improve practicality, they are less reliable compared to direct peer observation with the validated MOPET.

INTRODUCTION

The greatest factors impacting a medical learner's education are clinical teaching and role modeling, making effective methods for improving faculty teaching critical.[1,2] Validated assessment tools enhance clinical teaching by providing feedback to faculty on teaching skills.[3] Faculty assessment tools are limited by their application to specific clinical settings,[4] reliance on upward feedback,[5,6] use of perceptions rather than observations,[7-9] and lack of psychometric properties[10] or clear theoretical framework.

To fill the gap of a tool tailored to outpatient family medicine training, we created and initially validated the Mayo Outpatient Precepting Evaluation Tool (MOPET).[11] A drawback of the MOPET is the time required for peer observation, impeding practical use.

Modifying the MOPET for self-reflection could decrease reliance on peer observation. A well-developed self-reflection framework is Donald Schön's (1983) theory,[12] which includes retrospectively analyzing self and committing to future change.[13] Self-reflection is common in health care to promote independent growth;[13] therefore, it is a potentially helpful method for improving faculty teaching skills.

Additionally, resident-to-faculty feed back on clinical teaching is required by the Accreditation Council for Graduate Medical Education.[14] A resident-based feedback tool could help meet requirements while reducing reliance on peer observation.

Our study aimed to examine how self- and resident-reporting of teaching behaviors corresponded to the original peer-observation MOPET.

METHODS

This study was conducted at the Mayo Clinic Family Medicine Residency–Eau

Claire, Wisconsin, program. The project was approved by Mayo Clinic's Education Research Committee and deemed exempt by the Institutional Review Board. All residents (N = 15), core faculty, and community preceptors were invited to participate. Participation was voluntary, with no penalties or incentives. Resident-identifying information was collected only when residents were the observers.

The program's behavioral scientist, a faculty volunteer, and a resident used the original validated version of the MOPET (MOPET-Observer) to observe a preceptor for a half-day clinic (session). All observers were given standardized instructions for how to use the MOPET. As described previously, observers tallied how frequently a teaching behavior occurred over the course of a session.[11] A behavior could be counted only once per teaching encounter, defined as an instance of a resident entering the precepting space to discuss the care of a patient(s).

At the end of the session, the preceptor completed a modified MOPET (MOPET-Preceptor), asking whether specific teaching behaviors were used at least once during that session. Each precepted resident completed a modified MOPET (MOPET-Resident) at the end of the session, asking which precepting behaviors occurred during that session. All MOPET versions included identical items presented in the same order.

Observer data was dichotomized as "absent" if not observed or as "present" if the behavior was observed at least once during the session. Resident notation of preceptor behavior was rated as "present" if at least one resident observed the behavior.

Interrater reliability between observer and preceptor, observer and resident, and preceptor and resident was assessed via the $\varkappa$ (kappa) statistic (range −one to+one), which accounts for chance agreement.[15] A $\varkappa$ statistic can be calculated only if at least one observation is in each cell of the 2×2 contingency table. If variation is inadequate—for example, 100% prevalence—then a $\varkappa$ statistic cannot be calculated. Statistical analyses were performed via JMP version 18.0.1 (SAS Institute) with statistical significance set at P=0.05 without adjustment for multiple comparisons. Quantitative data lacked a normal distribution and is presented accordingly.

RESULTS

Across 48 sessions, 15 unique preceptors were observed two to four times each. The resident composition of sessions varied, with the number of first-, second- and third-year residents ranging from 0 to 3, with a median of 1 from each class. Sessions included five to 21 teaching encounters (IQR: 10–18; median: 15). The MOPET-Preceptor was completed for all but one session. Resident data were available for all sessions, coming from a single resident for two sessions, two residents for 16 sessions, three residents for 21 sessions, and four residents for nine sessions.

Table 1 shows how frequently teaching behaviors were observed or noted at least once during a session. Compared to the observers, residents more often reported experiencing 14

TABLE 1. Teaching Behavior Frequency During Half-Day Sessions Noted by Observer, Preceptor, and Precepted Residents

Teaching behavior	Observer count (range, IQR, median)	Observed prevalence (at least once during half-day session)	Preceptor noted prevalence (at least once during session)	Resident noted prevalence (at least once during session)
Encouraged resident to ask questions	0–13, 0–1, 0	33%	47%	100%
Did not interrupt	0–19, 6–13, 9.5	98%	94%	100%
Demonstrated interest	4–19, 8–15, 12	100%	100%	100%
Acknowledge resident teaching	0–5, 0–1, 0	44%	51%	83%
Available and ready	1–19, 7–13, 10	100%	98%	100%
Prioritized precepting	2–19, 7–15, 9	100%	100%	100%
Handled interruptions well	0–5, 0–2, 1	69%	72%	98%
Encouraged learning goal formation	0–5, 2–3, 2	92%	94%	98%
Discussed learning goal strategy	0–5, 0–2, 1	65%	64%	94%
Checked in on learning goal	0–4, 0–3, 1	67%	68%	96%
Asked questions	1–13, 3–6, 4	100%	94%	100%
Provided didactic teaching	1–13, 4–8, 6	100%	57%	94%
Used multimodal learning aids	0–11, 1–3, 1.5	77%	43%	58%
Asked about differential diagnosis	0–10, 1–4, 2	90%	94%	100%
Asked about alternative management	0–11, 0–3, 1	96%	77%	98%
Asked for rationale behind decision(s)	0–11, 0–3, 1	75%	74%	100%
Gave specific feedback	0–8, 1–3, 2	81%	72%	85%
Explained why assessment and plan appropriate (or not)	0–13, 1–4, 2.5	90%	89%	98%
Encouraged pursuing literature	0–9, 0–1, 0.5	50%	53%	85%
Reinforced self-directed learning	0–9, 0–1, 0	38%	40%	90%

of the 20 behaviors at least once in a session by an average of 24%. Residents reported lower prevalence of only two behaviors relative to observers: didactic teaching and use of multimodal learning aids. Preceptors self-reported using 10 behaviors less often and eight behaviors more often compared to observers. Residents reported more often experiencing 18 behaviors compared to preceptors' self-assessment.

Table 2 shows the agreement among observers, preceptors, and residents. We found statistically significant agreement between observers and preceptors for five of the 20 teaching behaviors ($\varkappa$ = 0.29 to 0.57); agreement between observers and preceptors could not be calculated for five behaviors due to 100% prevalence. Observers and residents significantly demonstrated agreement for 2 of the 20 behaviors measured ($\varkappa$ = 0.22 to 0.31), with nine behaviors lacking variation. Resident-preceptor agreement reached significance for three of the 20 behaviors ($\varkappa$ = 0.17 to 0.31), with agreement incalculable for eight behaviors.

DISCUSSION AND CONCLUSIONS

This study evaluated how well self- and resident-reported teaching behaviors corresponded to the validated, peer-observation–based MOPET. Among preceptors and observers, we found moderate agreement for discussing how to support a resident with their learning goal, checking in on the learning goal, and asking about a differential diagnosis. We found weak agreement among preceptors and observers for handling interruptions well and the use of multimodal learning aids. No significant agreement existed for half of the 20 teaching behaviors. We did not find evidence that preceptors tend to over- or underestimate their overall use of various teaching behaviors. In conclusion, self-assessment by preceptors showed limited agreement with observers using the validated MOPET.

Residents noted teaching behaviors more frequently than observers did. Resident assessment weakly corresponded with observers for preceptors' discussion of how to support the resident in their learning goal and use of multimodal learning aids. Residents' overreporting of positive teaching behaviors could be due to our method of data pooling from multiple residents for a single session and/or favorability bias. Residents may have avoided negatively portraying preceptors, despite assurance of anonymity. Our findings provide further evidence that power differentials influence resident-based assessments, which makes their ratings less reliable.[16]

TABLE 2. Interrater Reliability Among Observer, Preceptor, and Residents Reporting Occurrence of Teaching Behaviors at Least Once During a Half-Day Outpatient Clinic (N = 47)

Teaching behavior	Observer-preceptor (N = 47)	Observer-resident (N = 48)	Preceptor-resident (N = 47)
		$\varkappa$ (kappa) statistic	
Encouraged resident to ask questions	0.22		
Did not interrupt	−0.03		
Demonstrated interest			
Acknowledge resident teaching	0.28	0.04	0.09
Available and ready			
Prioritized precepting			
Handled interruptions well	**0.29**[a]	−0.04	−0.04
Encouraged learning goal	0.23	−0.03	−0.03
Discussed learning goal strategy	**0.54**[c]	**0.22**[a]	**0.21**[a]
Checked in on learning goal	**0.57**[d]	0.04	**0.17**[a]
Asked questions			
Provided didactic teaching			0.07
Used multimodal learning aids	**0.26**[a]	**0.31**[a]	0.26
Asked about differential diagnosis	**0.47**[b]		
Asked about alternative management	0.09	−0.03	−0.04
Asked for rationale behind decision(s)	−0.01		
Gave specific feedback	0.18	0.25	0.13
Explained why assessment and plan appropriate (or not)	0.10	−0.04	**0.31**[b]
Encouraged pursuing literature	0.19	0.04	−0.02
Reinforced self-directed learning	0.24	0.06	0.00

The $\varkappa$ (kappa) statistic is a measure of agreement between two dichotomous variables that accounts for chance (range −1 to 1). The blank cells indicate that a $\varkappa$ statistic could not be calculated due to a lack of variation—that is, at least one of the groups reported 100% prevalence of that behavior.

[a] $P \leq 0.05$

[b] $P \leq 0.01$

[c] $P \leq 0.001$

[d] $P \leq 0.0001$

https://doi.org/10.22454/FamMed.2026.238874

Our study found a high prevalence of multiple teaching behaviors such as demonstrating interest, being available to precept, and prioritizing precepting (98%–100%); the ceiling effect of preceptors consistently exhibiting these teaching behaviors limited our ability to calculate agreement. If other programs show that the same teaching behaviors are consistently prevalent, then an abbreviated MOPET excluding the most commonly observed behaviors may be more useful by reducing burden while maintaining validity.

This study is among the first to adapt a validated peer-observation tool for both self- and precepted resident reporting of teaching behaviors, allowing for direct comparison. Recall bias was limited due to data collection immediately following each session. This study examined occurrence of teaching behaviors rather than perceived teaching quality. Our single-site design limits generalizability.

Future studies replicating this work in other settings could further the MOPET's validation and generalizability. A revised MOPET could focus on less consistently used teaching behaviors to reduce ceiling effects. Overall, while self- and resident-reporting tools offer practical advantages, direct peer observation remains important for reliable, accurate assessment of teaching behaviors.

SUPPORT

This study was partially funded by the 2024 Endowment for Education Research Award from the Mayo Clinic College of Medicine and Science.

ACKNOWLEDGMENTS

The authors thank all those who supported and participated in this study, including faculty, preceptors, and residents.

REFERENCES

1. Bowen JL, Irby DM. Assessing quality and costs of education in the ambulatory setting: a review of the literature. *Acad Med*. 2002;77(7):621–680. doi:10.1097/00001888-200207000-00006
2. Royal KD. Quality teaching matters more than innovative curricula. *Am J Med*. 2017;130(4). doi:10.1016/j.amjmed.2016.11.014
3. Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. How reliable are assessments of clinical teaching? a review of the published instruments. *J Gen Intern Med*. 2004;19(9):971–977. doi:10.1111/j.1525-1497.2004.40066.x
4. Fluit CR, Bolhuis S, Grol R, Laan R, Wensing M. Assessing the quality of clinical teachers: a systematic review of content and quality of questionnaires for assessing clinical teachers. *J Gen Intern Med*. 2010;25(12):1337–1345. doi:10.1007/s11606-010-1458-y
5. Hinrichs L, Judd D, Hernandez M, Rapport M. Peer review of teaching to promote a culture of excellence: a scoping review. *JOPTE*. 2022;36(4):293–302. doi:10.1097/JTE.0000000000000242
6. Olvet DM, Willey JM, Bird JB, Rabin JM, Pearlman RE, Brenner J. Third year medical students impersonalize and hedge when providing negative upward feedback to clinical faculty. *Med Teach*. 2021;43(6):700–708. doi:10.1080/0142159X.2021.1892619
7. Castiglioni A, Shewchuk RM, Willett LL, Heudebert GR, Centor RM. A pilot study using nominal group technique to assess residents' perceptions of successful attending rounds. *J Gen Intern Med*. 2008;23(7):1060–1065. doi:10.1007/s11606-008-0668-z
8. Huff NG, Roy B, Estrada CA, et al. Teaching behaviors that define highest rated attending physicians: a study of the resident perspective. *Med Teach*. 2014;36(11):991–996. doi:10.3109/0142159X.2014.920952
9. Smith CA, Varkey AB, Evans AT, Reilly BM. Evaluating the performance of inpatient attending physicians: a new instrument for today's teaching hospitals. *J Gen Intern Med*. 2004;19(7):766–771. doi:10.1111/j.1525-1497.2004.30269.x
10. Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med*. 2005;20(12):1159–1164. doi:10.1111/j.1525-1497.2005.0258.x
11. Paulson D, Hidaka B, Nordin T. Creation and initial validation of the Mayo outpatient precepting evaluation tool. *Fam Med*. 2023;55(8):547–552. doi:10.22454/FamMed.2023.164770
12. Schön DA. *The Reflective Practitioner*. Basic Books; 1984.
13. Kinsella EA. Professional knowledge and the epistemology of reflective practice. *Nurs Philos*. 2010;11(1):3–14. doi:10.1111/j.1466-769X.2009.00428.x
14. Accreditation Council for Graduate Medical Education. ACGME program requirements for graduate medical education in family medicine. 2025. https://www.acgme.org/globalassets/pfassets/programrequirements/2025-reformatted-requirements/120_familymedicine_2025_reformatted.pdf
15. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 1960;20(1):37–46. doi:10.1177/001316446002000104
16. Beckman TJ, Lee MC, Mandrekar JN. A comparison of clinical teaching evaluations by resident and peer physicians. *Med Teach*. 2004;26(4):321–325. doi:10.1080/01421590410001678984