The Development of an Instrument for Faculty to Assess Resident-Led Large Group Teaching

Ariel S. Frey-Vogel, MD, MAT; Kristina Dzara, PhD, MMSc; Kimberly A. Gifford, MD; Erica Y. Chung, MD

From the Department of Pediatrics, MassGeneral Hospital for Children (AS Frey-Vogel and K Dzara), Boston, Mass; Department of Pediatrics, Dartmouth-Hitchcock Medical Center (KA Gifford), Lebanon, NH; Department of Pediatrics, Hasbro Children's Hospital (EY Chung), Providence, RI; Harvard Medical School (AS Frey-Vogel and K Dzara), Boston, Mass; The Warren Alpert Medical School of Brown University (EY Chung), Providence, RI; and Geisel School of Medicine at Dartmouth (KA Gifford), Hanover, NH
The authors have no conflicts of interest to disclose.
Address correspondence to Ariel S. Frey-Vogel, MD, MAT, Department of Pediatrics, MassGeneral Hospital for Children, 175 Cambridge St, fifth floor, Boston, MA 02114 (e-mail: afrey@mgh.harvard.edu, @AFrey_Vogel).
Received for publication July 19, 2019; accepted October 14, 2019.

ABSTRACT

OBJECTIVE: The Accreditation Council on Graduate Medical Education requires residents to teach and many residency programs assess resident teaching competency. While much formal resident-led teaching is for large groups, no corresponding published assessment instrument with validity evidence exists. We developed an instrument for faculty to assess pediatric resident-led large group teaching and gathered preliminary validity evidence.

METHODS: Literature review and our experience leading resident-as-teacher curricula informed initial instrument content. Resident focus groups from 3 northeastern pediatric residency programs provided stakeholder input. A modified Delphi panel of international experts provided iterative feedback. Three investigators piloted the instrument in 2018; each assessed 8 video recordings of resident-led teaching. We calculated Cronbach's alpha for internal consistency and intraclass correlation (ICC) for inter-rater reliability.

RESULTS: The instrument has 6 elements: learning climate, goals/objectives, content, promotion of understanding/retention, session management, and closure. Each element contains behavioral subelements. Cronbach's alpha was .844. ICC was excellent for 6 subelements, good for 1, fair for 1, and poor for 3.

CONCLUSIONS: We developed an instrument for faculty assessment of resident-led large group teaching. Pilot data showed assessed behaviors had good internal consistency, but inconsistent interrater reliability. With further development, this instrument has potential to assess resident teaching competency.

KEYWORDS: assessment instrument; instrument development; resident-as-teacher

ACADEMIC PEDIATRICS 2020;20:442−447

WHAT'S NEW

We describe the development and pilot study of an instrument for faculty to assess resident-led large group teaching. The instrument was developed to allow residents to receive structured actionable feedback and residency leadership to assess resident teaching competency.

RESIDENTS TEACH FREQUENTLY and desire more feedback on their teaching.[1] Furthermore, residency programs are required by the Accreditation Council for Graduate Medical Education to prepare residents to teach[2] and may benefit from data about resident teaching competency. According to Miller's pyramid of learner assessment, assessing actual resident teaching allows for assessment at the highest level of "does," which best predicts resident teaching competency.[3]

While residents often teach in the large group setting, there is a gap in the literature on assessment of large group teaching. Multiple assessment instruments have been developed to broadly assess teaching adult learners.[4−25] Of those designed for medical teaching, the instruments were largely designed for Observed Structured Clinical Examinations, 1:1 teaching, or medical student assessment of resident teaching. We identified no instruments with validity evidence for faculty evaluation of resident-led large group teaching. When programs assessed resident-led large group teaching in order to assess their resident as teacher (RAT) curricula, they utilized instruments without rigorous validity evidence.[19,26] There are recurring themes among published instruments, suggesting a shared understanding of key aspects of resident teaching, but no one instrument which uses the themes to allow for faculty assessment of resident-led large group teaching. These themes are: establishing learning climate, teacher

enthusiasm, initial learner assessment, goal communication, session control, instructional techniques, teacher knowledge, active learner involvement, learner evaluation and feedback, and application of concepts taught.

This study aims to systematically develop an instrument, building on these themes, for faculty to assess resident-led large group teaching. The instrument will serve 2 purposes, to provide: 1) residents objective assessment and feedback and 2) residency program directors data on individual resident teaching competency.

METHODS

STUDY SITES

The study authors are pediatricians (A.F.V., K.D., and E. C.) at 3 academic teaching hospitals who lead RAT curricula for pediatric residents and a medical education researcher (K.D.). These sites are MassGeneral Hospital for Children and Brown, both medium-sized urban residency programs, and Dartmouth, a smaller rural program, all in northeastern United States. At each site, all postgraduate year (PGY) 3 pediatric residents (and PGY4 medicine-pediatric residents at one site) participate in a RAT curriculum leading 2 to 8 case-based teaching conferences per year for a group of 10 or more mixed learners including medical students, residents, and faculty. Small groups require face-to-face interactions among and active involvement by all participants[27]; because these requirements are difficult to achieve with an audience of 10 or more, we define this as a large group. At each site, the specific requirements for conference vary, but most residents present a case. Residents receive varying degrees of mentorship, observation, and feedback on their teaching and the emphasis placed on large group teaching varies across sites. The Institutional Review Boards at Partners Healthcare, Dartmouth College, and Rhode Island Hospital exempted the instrument development from review and approved the pilot study.

INSTRUMENT DEVELOPMENT

We developed a list of core instrument content based on our experiences and a literature review (Figure, Step 1).[4−25] We revised the instrument with each subsequent round of input (2017−2018). First, a 1-hour focus group of 4 to 11 residents from all levels of training was held at each site to identify areas of assessment and feedback that would improve teaching. All residents were invited to participate via email; focus group sizes were determined by resident interest and availability. The focus groups were led by K.D., who has no evaluative role in any residency program (Figure, Step 2).

We then recruited a group of 14 international faculty with expertise in RAT curricula, assessment, and instrument development to create a modified Delphi panel; the study authors did not participate in this panel. The faculty were chosen based on our review of the RAT literature and personal knowledge of their work with a goal of recruiting a group with diverse geographical representation, medical specialty, training (MD and PhD), and expertise area. All panel members were asked for input simultaneously via survey for each round. Through 3 rounds of instrument review, the panel provided input on the importance of the subelements, the relationship between the subelements, and appropriate anchors for rating the behaviors (Figure, Steps 3−5).

PILOT STUDY

We invited all senior pediatric residents scheduled to lead upcoming conferences at each site to participate in a pilot (Spring 2018). Ten residents between the 3 sites volunteered to have 1 to 2 of their self-designed teaching sessions video recorded. Six of the teaching sessions were case-based discussions, 3 lessons learned from personal experiences, and 1 a gamified case series. Teaching sessions were 30 to 60 minutes depending on site requirements. Three study team members independently assessed each video using the instrument (A.F.V., E.C., and K.G.) without prior discussion about instrument use. We subsequently discussed our assessments of 2 videos together, revised the instrument to increase clarity and objectivity of the instrument's behaviors (Figure, Step 6 and Supplemental Content 1) and developed a guidebook for instrument use. We then independently reassessed the remaining 8 pilot teaching sessions using the guidebook.

DATA ANALYSIS

To determine internal consistency, Cronbach's alpha was calculated for the overall instrument as well as for each subelement. To determine inter-rater reliability, intraclass correlation (ICC) was calculated for each subelement by averaging the ICCs of each behavior within the subelement on the instrument. Calculations were completed using SPSS statistical package version 25 (SPSS Inc, Chicago, Ill).

RESULTS

INSTRUMENT DEVELOPMENT

The preliminary instrument encompassed a comprehensive list of potential content, including: 6 elements, 56 subelements, and an overall global rating scale (Figure, Step 1). From the resident focus group (Figure, Step 2), we found that the residents prefer comments over numerical scores because comments more easily translate to behavioral change. Residents indicated that the instrument would be helpful when preparing for teaching and act as a scaffold for debriefing with faculty. To meet resident needs, comment sections were added.

Fourteen Delphi panelists participated in round 1 (Figure, Step 3) to rate the subelements' importance with the goal of decreasing the number of subelements. The tool was subsequently revised based on the feedback. In round 2, the panelists determined which subelements were actually behaviors describing other subelements to further decrease the number of subelements. Eleven of the invited 14 Delphi panelists participated in this round (Figure, Step 4). In round 3, 8 of the invited 14 Delphi panelists (Figure, Step 5) provided input on instrument

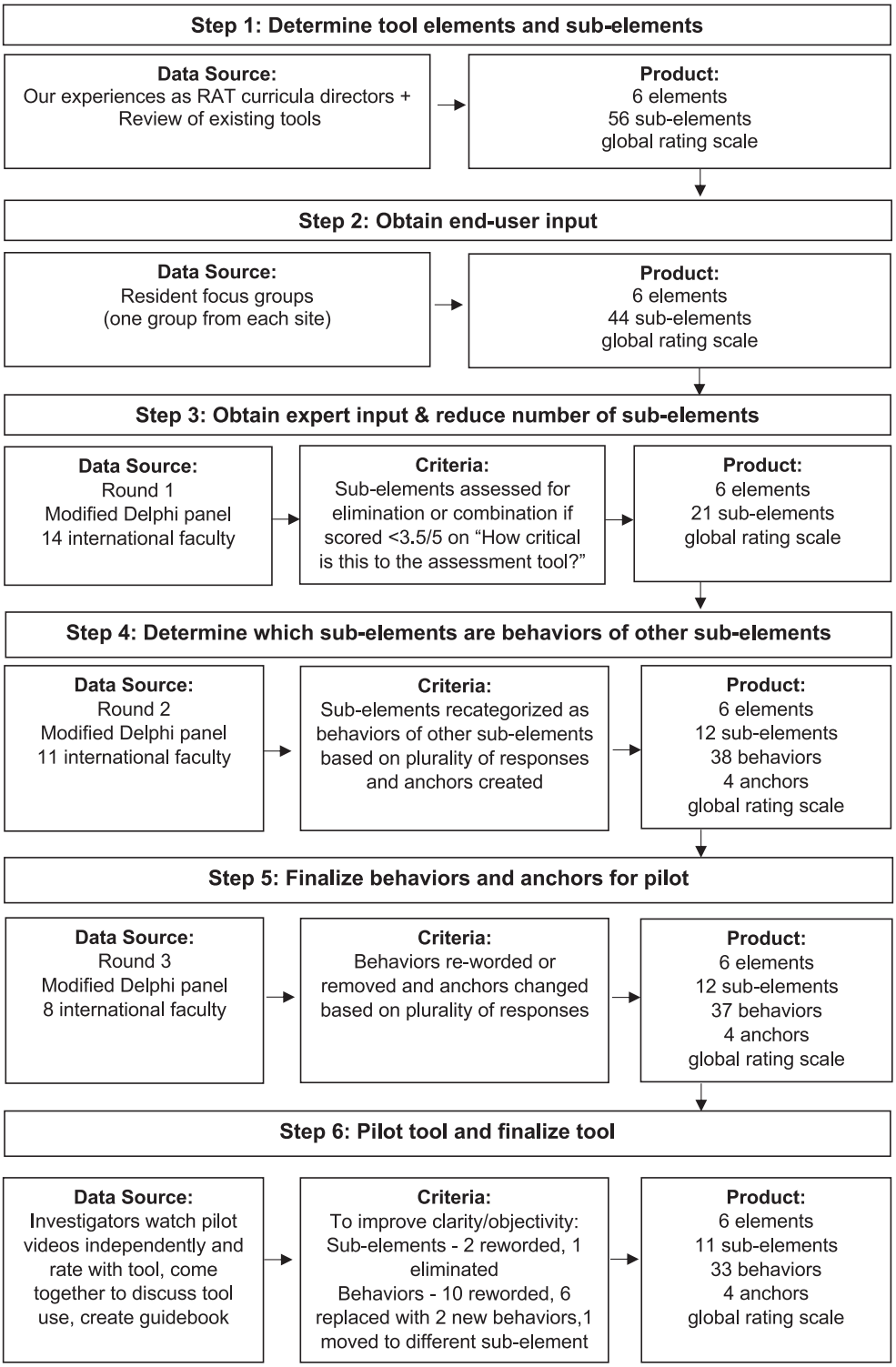

Figure. *Steps taken by the study authors to develop the assessment instrument during 2017 to 2018.*

structure (Supplemental Content 2). After this, the instrument had 6 elements, 12 subelements, 37 behaviors with anchors: "not at all," "partially," "consistently," and "as a role model" (Supplemental Content 3).

PILOT STUDY

After applying the instrument independently to 2 of the pilot videos, some behaviors were too subjective to allow for rater agreement. We eliminated 1 subelement and 4 behaviors, leaving a final tool with 6 elements, 11 subelements, and 33 behaviors. The remaining 8 teaching sessions taught by 6 different residents representing all sites were used for the pilot study. The physicians on our team (A.F.V., E.C., and K.G.) independently assessed all 8 sessions using the finalized instrument. From these 24 assessments of 8 teaching sessions, the overall internal consistency of the instrument was excellent (Cronbach's

Table. Intraclass Correlation of the Items on the Tool and Inter-rater Reliability of the 3 Raters Who Assessed 8 Resident Teaching Videos From All 3 Sites During the Pilot Study in 2018. Measures Calculated Using Cronbach's Alpha and Intraclass Correlation (ICC), Respectively, Where the Average of the ICC for All of the Behaviors Within a Subelement Made up the Subelement's ICC*

Element	Subelement	Cronbach's Alpha	ICC	ICC Interpretation
Learning climate	Created a respectful and open climate[a]		0.991	Excellent reliability
Learning climate	Clearly communicated the important of the topic and encouraged participant engagement throughout the presentation	.792	0.738	Good reliability
Goals and objectives	Set and communicated learner-centered, clear objectives appropriate for the time allotted	.943	0.800	Excellent reliability
Content of the talk	Demonstrated knowledge of the topic and used appropriate references	.376	0.886	Excellent reliability
Content of the talk	Tailored presentation level to participants' understanding of the material	.271	−0.345	Poor reliability
Promotion of understanding and retention	Explained concepts and interrelationships clearly	.616	0.783	Excellent reliability
Promotion of understanding and retention	Used effective questioning to promote learning and probed for supporting evidence or participants' thought processes	.777	0.929	Excellent reliability
Session management	Made efficient use of teaching time with appropriate pace and time spent on each objective[b]		0.496	Fair reliability
Session management	Content was logically organized with smooth transitions to optimize comprehension and retention	.828	−0.391	Poor reliability
Closure	Summarized key concepts and lessons learned[c]		0.378	Poor reliability
Closure	Explicitly encouraged further learning[d]		0.984	Excellent reliability

ICC indicates intraclass correlation.

There was no variability across learner or assessor for several behaviors and the internal consistency of the subelements to which they belonged was not calculated for this reason. Those behaviors were:

[a]Used respectful and inviting verbal and nonverbal language.

[b]Started and ended session on time.

[c]Determined if the learning objectives were met.

[d]Had participants articulate further learning goals for themselves on the topic.

*Categories of reliability derived from Cicchetti[28] where ICC <0.40 is poor reliability, 0.40 to 0.59 is fair reliability, 0.60 to 0.74 is good reliability, and 0.75 to 1.00 is excellent reliability.

alpha .844; Table). Four behaviors and their corresponding subelements were not analyzed because they showed no variance across assessor or teaching session. The ICC was excellent for 6 subelements, good for 1, fair for 1, and poor for 3 (Table).[28]

DISCUSSION

We created an instrument with potential for faculty to assess resident-led large group teaching. It was developed based on input from faculty who lead RAT curricula, literature review, resident focus groups, a modified Delphi process with RAT experts, and an instrument pilot. It incorporates themes from previously developed instruments in the literature[4–25] with the specific goal of allowing faculty to assess resident-led large group teaching, which was lacking in previously published tools. Further validation studies are needed for use in high-stakes assessment; the instrument provides a framework to help residents plan their teaching and guide feedback.

When utilized in the pilot study, the instrument had overall high internal consistency. Some subelements had lower internal consistency or were not included in the calculations in part due to the lack of variability in scores across assessor and observed teaching session. For example, the section on closure included several behaviors which no residents performed in the pilot ("determine if learning objectives were met" and "had participants articulate further learning goals for themselves on the topic"). Conversely, other behaviors ("used respectful and inviting verbal and nonverbal language" and "started and ended session on time") were not included in the calculations because they were always accomplished by the residents in the pilot study. These behaviors should be assessed because they are fundamental aspects of teaching, yet their inclusion makes calculating internal consistency difficult. These issues will be assessed in future validity studies. We did not include qualitative feedback in the pilot because its purpose is as formative feedback which was not our purpose with the pilot.

Inter-rater reliability was excellent or good for 7 of the 11 subelements. The low inter-rater reliability for the other 4 elements may be due to: 1) the assessors not undergoing a formal process to develop a shared mental model for instrument use and 2) some instrument behaviors having little variability among assessors or teaching episodes. One instrument with rigorous validity evidence assesses faculty-led large group teaching[13,29]; while initially low,[13] the ICC improved with creating a shared mental model and behavioral descriptors.[29] Subsequently, novice assessor reliability improved through frame-of-reference (FOR) training, where experts explained their

shared mental model and used it to give novices feedback.[30] FOR training would be a useful technique in training assessors for our instrument.

Our study has several limitations. The assessors developed the instrument, which may have inflated the instruments' internal consistency and ICC in the pilot study. The instrument was not tested outside of academic pediatric settings. However, the teaching behaviors assessed should be applicable to any resident-led large group teaching setting, as evidenced by the recurring themes found in teaching assessment instruments across disciplines.[4–25] The instrument was only piloted on senior residents, who may have less variation in teaching competency than all residents. We also did not conduct a formal qualitative analysis of the focus group data. Furthermore, we did not follow a strict Delphi panel method, but rather summarized the experts' perceptions. We had a reduction in the number of Delphi panel participants with each round which was likely due to the degree of labor required to review the large number of behaviors and work spanning several months.

Our next step is to develop a shared mental model and update our instrument guidebook. Subsequently, we will use a FOR approach to train faculty at our 3 sites. These faculty will assess teaching sessions recorded over the course of a full academic year to determine the instrument's: ICC, internal consistency, and agreement with clinical competency committee assessment of senior residents on the nonreported ACGME pediatric teaching milestone, "Develop the necessary skills to be an effective teacher."[31] Because the instrument remains lengthy, we hope to determine which behaviors are duplicative and could be removed using factor analysis. After refining the tool, we will consider whether a shorter version of the tool could assess less formal teaching, such as on rounds. It will be important to examine resident perceptions of the instrument after its use.

In conclusion, we created an instrument for faculty to assess resident-led large group teaching. Our pilot data demonstrate that our instrument has good internal consistency when used by instrument developers. Without formal rater training, our inter-rater reliability was excellent for over half of the instruments' subelements. With formal rater training, our instrument may enable faculty to assess resident teaching at Miller's highest level and serve as a tool to improve their ability to give feedback on resident teaching and our residents to have a guide for planning their teaching.

Acknowledgments

The authors wish to thank Virginia Reed, PhD, MS, for her statistical support, the residents who volunteered to participate in our focus groups and to be video recorded for the pilot study, the experts who participated in our modified Delphi panel, the chief residents who video recorded the resident teaching sessions, and our residency program directors for their support. We would also like to thank Jacob Johnson, MD, and Daniel Saddowi-Konefka, MD, for their feedback on an earlier version of this work.

Financial statement: This study was funded by an Association of American Medical Colleges' Northeastern Group on Educational Affairs collaborative research grant. The PI for this grant was Ariel Frey-Vogel, MD, MAT. The funder had no role in study design, data analysis, collection, or interpretation, in the writing of the manuscript, or in the decision to submit the manuscript for publication.

Supplementary Data

Supplementary data related to this article can be found online at https://doi.org/10.1016/j.acap.2019.10.010.

References

1. Tuck KK, Murchison C, Flores C, et al. Survey of residents' attitudes and awareness toward teaching and student feedback. *J Grad Med Educ*. 2014;6:698–703.
2. Accreditation Council for Graduate Medical Education. ACGME Common Program Requirements. Section IV.A.5.c)(8) https://www.acgme.org/Portals/0/PFAssets/ProgramRequirements/CPRs_2017-07-01.pdf. Published 2017. Accessed January 10, 2019.
3. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med*. 1990;65:S63–S67.
4. Morrison EH, Boker JR, Hollingshead J, et al. Reliability and validity of an objective structured teaching examination for generalist resident teachers. *Acad Med*. 2002;77(August 2001):S29–S32.
5. Ricciotti HA, Dodge LE, Head J, et al. A novel resident-as-teacher training program to improve and evaluate obstetrics and gynecology resident teaching skills. *Med Teach*. 2012;34:e52–e57.
6. Busari JO, Scherpbier AJJA, Van Der Vleuten CPM, et al. A two-day teacher-training programme for medical residents: investigating the impact on teaching ability. *Adv Health Sci Educ Theory Pract*. 2006;11:133–144.
7. James MT, Mintz MJ, Mclaughlin K. Evaluation of a multifaced "resident-as-teacher" educational intervention to improve morning report. *BMC Med Educ*. 2006;6:20.
8. Johnson NR, Chen J. Medical student evaluation of teaching quality between obstetrics and gynecology residents and faculty as clinical preceptors in ambulatory gynecology. *Am J Obstet Gynecol*. 2006;195:1479–1483.
9. Gaba ND, Blatt B, Macri CJ, et al. Improving teaching skills in obstetrics and gynecology residents: evaluation of a residents-as-teachers program. *Am J Obstet Gynecol*. 2007;196:87.e1–87.e7.
10. Spooren P, Mortelmans D, Denekens J. Student evaluation of teaching quality in higher education: development of an instrument based on 10 Likert-scales. *Assess Eval High Educ*. 2007;32:667–679.
11. Trujillo JM, DiVall MV, Barr J, et al. Development of a peer teaching-assessment program and a peer observation and evaluation tool. *Am J Pharm Educ*. 2008;72:147.
12. Newman L, Tibbles CD, Atkins KM, et al. Resident-as-teacher DVD series. *MedEdPORTAL*. 2015;11:10152.
13. Newman LR, Lown BA, Jones RN, et al. Developing a peer assessment of lecturing instrument: lessons learned. *Acad Med*. 2009;84:1104–1110.
14. Ilgen JS, Takayesu JK, Bhatia K, et al. Back to the bedside: the 8-year evolution of a resident-as-teacher rotation. *J Emerg Med*. 2011;41:190–195.
15. Keller JM, Blatt B, Plack M, et al. Using a commercially available web-based evaluation system to enhance residents' teaching. *J Grad Med Educ*. 2012;4:64–67.
16. Zackoff M, Jerardi K, Unaka N, et al. An observed structured teaching evaluation demonstrates the impact of a resident-as-teacher curriculum on teaching competency. *Hosp Pediatr*. 2015;5:342–347.
17. Pettit JE, Axelson RD, Ferguson KJ, et al. Assessing effective teaching. *Acad Med*. 2015;90:94–99.
18. Knight C, Windish D, Haist S, et al. The SGIM TEACH program: a curriculum for teachers of clinical medicine. *J Gen Intern Med*. 2017;32:948–952.
19. D'Eon MF. Evaluation of a teaching workshop for residents at the University of Saskatchewan: a pilot study. *Acad Med*. 2004;79:791–797.

20. Irby D, Rakestraw P. Evaluating clinical teaching in medicine. *J Med Educ*. 1981;56:181–186.
21. Litzelman DK, Stratos GA, Marriott DJ, et al. Factorial validation of a widely disseminated educational framework for evaluating clinical teachers. *Acad Med*. 1998;73:688–695.
22. Copeland HL, Hewson MG. Developing and testing an instrument to measure the effectiveness of clinical teaching in an academic medical center. *Acad Med*. 2000;75:161–166.
23. Furney SL, Orsini AN, Orsetti KE, et al. Teaching the one-minute preceptor: a randomized controlled trial. *J Gen Intern Med*. 2001;16:620–624.
24. Zabar S, Hanley K, Stevens DL, et al. Measuring the competence of residents as teachers. *J Gen Intern Med*. 2004;19:530–533.
25. van der Hem-Stokroos HH, van der Vleuten CPM, Daelmans HEM, et al. Reliability of the clinical teaching effectiveness instrument. *Med Educ*. 2005;39:904–910.
26. Post RE, Quattlebaum RG, Benich JJ. Residents-as-teachers curricula: a critical review. *Acad Med*. 2009;84:374–380.
27. Walton H. Small group methods in medical teaching. *Med Educ*. 1997;31:459–464.
28. Cicchetti D. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instrument in psychology. *Psychol Assess*. 1994;6:284–290.
29. Newman LR, Brodsky DD, Roberts DH, et al. Developing expert-derived rating standards for the peer assessment of lectures. *Acad Med*. 2012;87:356–363.
30. Newman LR, Brodsky D, Jones RN, et al. Frame-of-reference training: establishing reliable assessment of teaching effectiveness. *J Contin Educ Health Prof*. 2016;36:206–210.
31. American Board of Pediatrics and Accreditation Council for Graduate Medical Education. The Pediatrics Milestone Project. 201259.