Leadership Observation and Feedback Tool: A Novel Instrument for Assessment of Clinical Leadership Skills

Sandra K. Oza, MD, MA
Sandrijn van Schaik, MD, PhD
Christy K. Boscardin, PhD
Read Pierce, MD
Edna Miao, MD
Tai Lockspeiser, MD, MHPE
Darlene Tad-y, MD
Eva Aagaard, MD
Anda K. Kuo, MD

ABSTRACT

Background While leadership training is increasingly incorporated into residency education, existing assessment tools to provide feedback on leadership skills are only applicable in limited contexts.

Objective We developed an instrument, the Leadership Observation and Feedback Tool (LOFT), for assessing clinical leadership.

Methods We used an iterative process to develop the tool, beginning with adapting the Leadership Practices Inventory to create an open-ended survey for identification of clinical leadership behaviors. We presented these to leadership experts who defined essential behaviors through a modified Delphi approach. In May 2014 we tested the resulting 29-item tool among residents in the internal medicine and pediatrics departments at 2 academic medical centers. We analyzed instrument performance using Cronbach's alpha, interrater reliability using intraclass correlation coefficients (ICCs), and item performance using linear-by-linear test comparisons of responses by postgraduate year, site, and specialty.

Results A total of 377 (of 526, 72%) team members completed the LOFT for 95 (of 519, 18%) residents. Overall ratings were high—only 14% scored at the novice level. Cronbach's alpha was 0.79, and the ICC ranged from 0.20 to 0.79. Linear-by-linear test comparisons revealed significant differences between postgraduate year groups for some items, but no significant differences by site or specialty. Acceptability and usefulness ratings by respondents were high.

Conclusions Despite a rigorous approach to instrument design, we were unable to collect convincing validity evidence for our instrument. The tool may still have some usefulness for providing formative feedback to residents on their clinical leadership skills.

Introduction

There is increasing recognition that physicians need to be prepared to lead health care teams, and leadership increasingly is included in residency education.[1–4] Yet residents rarely receive feedback to aid leadership skills development, and tools to assess these skills are limited.[5,6] Most tools were developed for specific, high-stakes clinical situations, such as resuscitations and crises in the operating room.[7–14] Leadership skills for these high-intensity, time-limited situations are different from those required for the longitudinal context of ambulatory clinics and inpatient units.

The Leadership Practice Inventory (LPI) is an instrument with validity evidence in the business literature.[15] The LPI has been used for several years to provide multisource feedback to pediatrics residents in a leadership track at the University of California, San Francisco (UCSF),[2] but it has limitations in usability due to its length and perceived lower applicability to health care. We sought to adapt the LPI into a shorter, clinically relevant instrument to guide feedback for residents on their leadership skills.

DOI: http://dx.doi.org/10.4300/JGME-D-18-00113.1

Editor's Note: The online version of this article contains the 10-item Leadership Observation and Feedback Tool instrument; participating residents by program, site, and postgraduate year; and raters by professional role.

Methods

We developed the Leadership Observation and Feedback Tool (LOFT) using an iterative, mixed-methods approach. We collected validity evidence by applying a unitary view of validity as described by Messick,[16] focusing on content validity, response process, internal structure, and relationship to other variables.[17]

Participants and Settings

We conducted this study in May 2014 among residents in internal medicine (IM) and pediatrics at 2 large academic centers: UCSF and the University of Colorado School of Medicine (CUSOM).

Instrument Development

First, we created an open-ended survey instrument (TABLE 1) based on the 5 domains of the LPI to identify behaviors that constitute clinical leadership, and we collected information about team leadership behaviors for 20 residents on inpatient rotations at UCSF (13 IM and 7 pediatrics) from 86 team members of different professional backgrounds working with those residents on the inpatient unit. Two investigators (S.v.S. and E.M.) independently coded 5 randomly selected survey instruments for IM residents, discussed and reconciled differences, and created a preliminary coding scheme using a thematic approach.[18] They repeated this process with the next 5 instruments, made refinements to the coding scheme, and subsequently coded all IM instruments, organizing the data into themes and subthemes. They repeated this process for the instruments for the 7 pediatrics residents, compared the theme list to the IM themes, and created a combined theme list. Three investigators (S.v.S., R.P., and A.K.K.) then reviewed the theme list for internal consistency and coherency and translated it into a list of practices and characteristics organized within themes of clinical leadership.

We used HyperRESEARCH 3.0 software (ResearchWare Inc, Randolph, MA) to organize and analyze the qualitative data. We identified 30 clinical leadership behaviors distributed over 10 themes (TABLE 2). To collect evidence for content validity, we asked 15 international experts in health care teamwork and leadership from a variety of professions to review the list of behaviors, using a modified Delphi approach.[19] In the first round, experts indicated the importance of each behavior to clinical team leadership on a 4-point Likert scale (*strongly disagree* to *strongly agree*). We used a content validity index to quantify agreement between raters and found high levels of agreement for all 30 clinical leadership behaviors (≥ 0.80 for all).[20] Experts also suggested refinements to the list and 7 additional clinical leadership behaviors. In a second round, the experts identified the developmental stage at which the 37 behaviors would be exhibited (from novice to expert). We then labeled each behavior with the developmental stage suggested by the largest number of experts (10 at the novice stage, 10 at the advanced beginner stage, and 17 at the proficient stage). Based on this, we constructed a 10-item instrument, with 3 unique behavioral anchors on a 5-point developmental scale (provided as online supplemental material). We pilot tested this 10-item instrument (LOFT) with a new cohort of 78 team members (of 20 IM and pediatrics residents at UCSF). The average rating per item ranged from 4.45 to 4.73 (5-point scale). Because of these uniformly high ratings, we consulted with local assessment experts and revised the instrument to avoid overrating and "halo" effects. The revisions consisted of (1) removing numerical values associated with behavioral anchors; (2) incorporating reverse-scored items; and (3) breaking up items with compound behaviors. The final instrument consisted of 29 items within the original 10 themes and focused on observable behaviors, each with 3 descriptors (TABLE 2).

What was known and gap
Residents benefit from feedback on their evolving clinical leadership skills, yet there is a dearth of validated instruments.

What is new
A study sought to adapt the domains of a validated leadership assessment to residents' everyday clinical context, and to assess the resulting instrument for validity evidence.

Limitations
Limited specialty and institutional context limit generalizability; there is a potential for response and social desirability bias.

Bottom line
This relatively important negative study of a leadership tool was not able to provide validity evidence for its use in rating residents' clinical leadership skills.

Instrument Testing

We invited IM and pediatrics residents in all postgraduate years (PGYs) to identify at least 5 clinical team members who could give feedback on their clinical leadership and asked team members via e-mail to complete the LOFT in an online platform (SurveyMonkey, San Mateo, CA). We included 5 questions to solicit feedback on the LOFT to collect evidence for response process and usability of the instrument. We also collected evidence for response process by examining the frequency of "not applicable" (N/A) ratings.

The UCSF and the CUSOM Multiple Institutional Review Boards approved the study.

Data Analysis

We calculated the frequencies of responses for each item and descriptive statistics for the items inviting feedback on the LOFT. We calculated Cronbach's alpha for the instrument overall and within the 10 themes for evidence of the instrument's internal structure, and we calculated intraclass correlation coefficients (ICCs) for interrater reliability for each PGY level. We then performed linear-by-linear association tests to compare ratings residents

TABLE 1
Leadership Practice Inventory (LPI) Domains and Questions in Pre–Leadership Observation and Feedback Tool (LOFT)

LPI Domains	Pre-LOFT Questions
Model the Way: A leader sets clear, convincing examples of the way people should be treated, how goals should be pursued, and what standards count. Through thoughtful action, leaders help others to succeed.	Q1: How did the resident model high-quality teamwork and leadership?
	Q2: How did the resident set expectations and ensure clear communication among team members?
	Q3: How could the resident communicate more clearly and effectively?
Inspire a Shared Vision: A leader utilizes charisma, passion, and persuasion to excite others about the future. Leaders convince people to embrace their visions of excellence.	Q4: How did the resident motivate team members?
	Q5: How did the resident establish common goals?
	Q6: How could the resident increase team members' belief in their work and common goals?
Challenge the Process: A leader seeks innovative ways to improve organizations, even when doing so involves risk. In pursuit of a better way, leaders accept mistakes and frame failures as learning opportunities.	Q7: How did the resident handle challenges that the team encountered?
	Q8: In what ways did the resident help team members learn from mistakes?
	Q9: How could the resident more effectively encourage the team or individual members to improve performance?
Enable Others to Act: A leader fosters collaboration and morale by emphasizing mutual respect, trust, and dignity. Leaders use and stretch the unique capacities of individuals and teams, increasing performance.	Q10: How did the resident create a supportive and respectful team environment?
	Q11: In what ways did the resident coordinate task distribution according to team members' skills and abilities?
	Q12: How could the resident enhance collaboration within the team while also encouraging individual effort?
Encourage the Heart: A leader celebrates team member contributions and successes to show appreciation for determination, dedication, and hard work.	Q13: In what ways did the resident display appreciation for team members' work?
	Q14: How did the resident express confidence in team members and celebrate successes?
	Q15: What else could the resident do to create a culture of appreciation?

received, grouped by PGY level, site, and specialty as evidence of relationship to other variables. We compared the ratings of physician evaluators (attendings, fellows, and residents) with those of other team members (nurses, pharmacists, medical students, and allied health professionals). Finally, we calculated the total instrument score for each resident, determined the mean instrument score for each PGY level and compared them using analysis of variance to provide evidence for the developmental nature of the construct, and calculated the effect size using Cohen *d*. We set statistical significance at $P = .05$ and used SPSS version 23 (IBM Corp, Armonk, NY) for all statistical calculations.

Results

Descriptive Statistics and Evidence of Internal Structure

Of 519 invited, 95 residents (18%) participated and identified 526 team members to complete the instrument. Of those, 377 (72%) accessed the survey, including 69 (18%) from professions outside medicine (details provided in online supplemental material). Three respondents accessed the survey but did not answer any items, and they were excluded from analysis. Cronbach's alpha for internal consistency was 0.79, indicating high item reliability. Cronbach's alpha within the original 10 themes ranged from 0.20 (effectively handles challenging situations) to 0.76 (shows appreciation to motivate the team). The ICC for items for each PGY level ranged from 0.20 to 0.79, with only 5 items having an ICC greater than 0.60 (items 1, 2, 6, 19, and 22; TABLE 2). Across all items, the majority of residents received ratings that we postulated to be consistent with proficient leadership skills (between 54% and 63% depending on PGY year; TABLE 3), and only a small proportion (13%–14%) received ratings consistent with novice leadership skills.

Evidence of Relationship to Other Variables

Comparing residents' performance by group using a linear-by-linear test revealed a few differences at the

TABLE 2
Themes and Associated Behaviors From Pre–Leadership Observation and Feedback Tool (LOFT) Qualitative Data Analysis and 29-Item LOFT

Themes and Associated Behaviors	LOFT Item	Item Response Options			
1. Shows appreciation to motivate the team ▪ Thanks team members for their work ▪ Gives praise for work well done ▪ Acknowledges/highlights successes and accomplishments ▪ Does things for the team to show appreciation (eg, brings food)	1	N/A	Does not praise the team or team members	Often praises the team or team members	Consistently highlights team successes and praises the team
	2	N/A	Does not do things for the team that demonstrate appreciation	Often does things for the team that demonstrate appreciation	Consistently celebrates team successes
2. Balances autonomy and supervision ▪ Promotes ownership—allows team members to independently generate and execute plans ▪ Balances allowing independent work with appropriate supervision ▪ Is confident in other team members' work	3	N/A	Does not micromanage team members	Sometimes micromanages team members	Consistently micromanages team members
	4	N/A	Gives team members little room to independently generate and execute plans, even when they are competent to do so	Often gives competent team members room to independently generate and execute plans	Consistently allows competent team members to independently generate and execute plans
	5	N/A	Does not provide adequate supervision for team members	Often provides adequate supervision for team members	Consistently provides adequate supervision for team members
3. Is accessible and involved ▪ Is available and approachable ▪ Checks in with team members frequently	6	N/A	Does not take an interest in team members	Often takes an interest in team members	Consistently invests in relationships with team members
	7	N/A	Is never available	Is inconsistently available	Is consistently available
	8	N/A	Is often dismissive and difficult to approach	Can be dismissive at times but generally easy to approach	Is always easy to approach and never dismissive
4. Ensures collaboration with team members for shared decision-making ▪ Solicits input from all members of the team, including nonphysicians ▪ Listens carefully to others ▪ Communicates directly and clearly with all team members ▪ Promotes mutual goal-setting and shared decision-making	9	N/A	Does not listen to suggestions or concerns of team members	Often listens to suggestions and concerns of team members	Consistently listens to suggestions and concerns of team members
	10	N/A	Does not solicit input from team members	Solicits input from team members inconsistently	Consistently solicits input across members of the team
	11	N/A	Does not check in with team members	Inconsistently checks in with team members	Consistently checks in with team members
	12	N/A	Makes timely and firm decisions	Can sometimes be indecisive	Consistently struggles with decision-making
	13	N/A	Does not engage in collaborative decision-making	Includes some team members in collaborative decision-making	Consistently engages team members across professions in collaborative decision-making

TABLE 2
Themes and Associated Behaviors From Pre–Leadership Observation and Feedback Tool (LOFT) Qualitative Data Analysis and 29-Item LOFT (continued)

Themes and Associated Behaviors	LOFT Item	Item Response Options			
5. Assists with workload management ▪ Distributes work appropriately and fairly based on skill level ▪ Helps with any tasks, particularly at busy times ▪ Incorporates individual learning needs when delegating tasks	14	N/A	Does not help out when the team's workload is high	Often helps out when the team's workload is high	Consistently helps out when the team's workload is high
	15	N/A	Does not prioritize tasks for the team	Often prioritizes tasks for the team	Consistently prioritizes tasks for the team
	16	N/A	Does not distribute the workload among team members	Distributes the workload among team members but not always appropriately or fairly	Consistently distributes the workload appropriately and fairly among team members
6. Provides supportive feedback ▪ Provides specific and constructive feedback, identifies areas for improvement ▪ Provides positive feedback and encouragement ▪ Gives feedback frequently	17	N/A	Does not provide feedback to team members	Often provides feedback to team members, but not always specific, balanced, or timely	Consistently provides specific, balanced, and timely feedback to team members
7. Effectively handles challenging situations ▪ Faces challenges through application of problem-solving skills ▪ Has the ability to be assertive ▪ Has a positive attitude, even during difficult times ▪ Stays calm in stressful situations	18	N/A	Always appears calm in challenging situations	Sometimes exhibits stress in challenging situations	Consistently exhibits stress in challenging situations
	19	N/A	Never avoids challenging situations	Often faces challenging situations	Consistently avoids challenging situations
	20	N/A	Has a negative attitude	Attitude is neither negative nor positive	Has a positive attitude
	21	N/A	Does not manage conflict effectively	Often manages conflict effectively	Consistently handles conflict effectively
8. Promotes a learning environment ▪ Creates an environment in which team members can discuss and learn from mistakes ▪ Places an emphasis on teaching and learning	22	N/A	Does not pay attention to the individual learning needs of team members	Often takes time to explore the individual learning needs of team members	Consistently explores individual learning needs of team members
	23	N/A	Consistently more focused on learning than on completing tasks	Balances team learning with task completion	Consistently more focused on completing tasks than learning

TABLE 2
Themes and Associated Behaviors From Pre–Leadership Observation and Feedback Tool (LOFT) Qualitative Data Analysis and 29-Item LOFT (continued)

Themes and Associated Behaviors	LOFT Item	Item Response Options			
9. Models professional behavior ▪ Models how to treat others (respectful to staff and patients, caring toward patients) ▪ Models dedication to and passion for high-quality patient care	24	N/A	Does not take responsibility for mistakes	Often takes responsibility for mistakes but does not always show effort toward self-improvement	Consistently takes responsibility for mistakes and models self-improvement
	25	N/A	Does not express awareness of strengths and weaknesses	Inconsistently expresses awareness of strengths and weaknesses	Consistently expresses awareness of strengths and weaknesses
	26	N/A	Never shows disrespectful behavior toward others	Sometimes disrespectful toward others	Consistently disrespectful toward others
10. Establishes expectations and goals ▪ Sets clear expectations and goals at the beginning ▪ Frequently reminds others of goals/ expectations ▪ Ensures that expectations and goals are achieved	27	N/A	Does not set expectations of team members	Explains expectations to team members but does not verify understanding	Consistently ensures that team members understand expectations
	28	N/A	Does not engage in development of shared goals for the team	Inconsistently engages in development of shared goals for the team	Consistently engages in development of shared goals for the team
	29	N/A	Does not hold team members accountable for achieving goals and meeting expectations	Inconsistently holds team members accountable for achieving goals and meeting expectations	Consistently holds team members accountable for achieving goals and meeting expectations

TABLE 3
Distribution of Ratings per Leadership Observation and Feedback Tool Item

Item	PGY-1, %				PGY-2, %				PGY-3, %				*P* Value	
	N/A	NV	AB	PRF	N/A	NV	AB	PRF	N/A	NV	AB	PRF	PGY[a]	N/A[b]
1	5	1	51	43	4	0	27	69	4	7	33	57	.044[c]	.49
2	3	3	45	49	2	0	42	56	4	4	36	57	.27	.99
3	24	67	8	1	4	71	23	2	7	63	29	1	.003[c]	< .001[c]
4	24	0	26	50	4	1	29	66	4	1	25	70	.51	< .001[c]
5	30	...	16	54	9	...	14	77	7	...	13	80	.08	< .001[c]
6	1	0	19	80	0	1	16	83	7	0	21	72	.65	.009[c]
7	0	...	1	99	1	...	3	96	0	...	3	97	.43	.88
8	...	...	5	95	...	...	8	92	...	...	4	96	.91	N/A
9	1	...	18	80	0	...	16	84	3	...	18	79	.93	.65
10	3	0	16	81	2	0	10	88	4	1	16	79	.97	.75
11	3	0	3	94	2	1	2	95	7	0	3	91	.89	.42
12	5	72	22	1	1	75	22	3	3	79	18	0	.42	.19
13	3	1	17	80	0	0	17	83	3	0	15	83	.38	.63
14	7	0	13	80	4	1	14	81	7	3	11	80	.62	.56
15	16	1	26	58	6	1	18	75	4	0	28	68	.39	.001[c]
16	44	...	4	52	12	...	4	83	15	...	4	82	.49	< .001[c]
17	31	2	20	47	15	1	21	63	15	3	17	66	.24	< .001[c]
18	1	66	33	1	1	70	23	6	0	71	24	5	.95	.73
19	3	66	30	1	4	71	20	4	7	66	26	1	.67	.37
20	...	...	1	99	...	...	5	95	...	...	8	92	.014[c]	N/A
21	19	...	31	50	17	...	23	60	20	...	17	63	.011[c]	.98
22	29	0	34	36	15	1	28	57	12	0	29	59	.027[c]	< .001[c]
23	7	1	89	4	5	3	86	6	7	5	87	1	.07	.76
24	17	0	5	78	14	0	7	79	18	1	4	76	.63	.63
25	11	3	11	76	12	0	10	78	13	1	4	82	.06	.75
26	0	98	1	1	0	95	3	2	3	93	0	4	.15	.037[c]
27	48	3	9	40	17	1	6	75	17	0	12	71	.06	< .001[c]
28	23	1	5	72	9	1	9	82	13	0	9	78	.49	.007[c]
29	46	0	3	51	15	1	7	77	18	0	8	74	.30	< .001[c]
Mean	14	13	19	54	6	14	18	63	8	14	18	61	...	...

Abbreviations: PGY, postgraduate year; N/A, not applicable; NV, novice; AB, advanced beginner; PRF, proficient.
[a] This column presents the *P* value for linear-by-linear association tests comparing proportions of item responses across PGY-1 to PGY-3.
[b] This column presents the *P* value for linear-by-linear association tests between proportions of "not applicable" (N/A) item responses compared to all other responses.
[c] $P \leq .05$.

item level (TABLE 3). For 5 items, there were significant differences between PGY groups; senior residents received ratings consistent with more advanced leadership than PGY-1 residents on 3 items (1, 21, and 22), and for 2 items (3 and 20) the reverse was true. The comparison of total scores on the overall LOFT leadership measure among PGY groups revealed that, as hypothesized, PGY-1 scores (mean = 68, SD = 13) were significantly lower compared with PGY-2 (mean = 75; SD = 12; $d = 0.56$; $P < .001$ for comparison) and PGY-3 (mean = 75; SD = 14; $d = 0.52$; $P < .001$ for comparison) scores. For 6 items (3, 12, 17, 21, 28, and 29) performance ratings from physicians (residents, fellows, and attending physicians) were lower than from other raters.

Evidence of Response Process

The percentage of N/A responses ranged from zero (items 8 and 20) to 29.7% (item 27), with a mean percentage of 9.5% across all items. Linear-by-linear test comparisons of the frequency of N/A responses by PGY revealed significant differences for 12 items (3–6, 15–17, 22, and 26–29). For all but 2 of these items (6, 26), N/A was more likely to be given for a

PGY-1 than for a PGY-2 or a PGY-3 resident. The majority of evaluators agreed or strongly agreed the instrument is easy to use (88%, 323 of 367), useful for providing feedback on leadership skills (72%, 261 of 365), and provides an opportunity to give feedback on skills that are not currently included in feedback to residents (73%, 264 of 361).

Discussion

During pilot testing of the LOFT, team members from multiple professions rated its acceptability and utility highly. Overall scores showed significant differences between interns (PGY-1) and more senior residents (PGY-2 or PGY-3), which provides some validity evidence for relationship to other variables, which was not found elsewhere in our analyses. Further, ratings for most items on the LOFT were high across all PGY years and did not discriminate between residents at different levels. Our ability to collect further validity evidence based on the internal structure of LOFT was limited by these high ratings.[16,17]

With both iterations of the LOFT, we saw a ceiling effect, with residents across training years receiving ratings consistent with proficient leadership behaviors. We considered several explanations for these findings. We rejected the explanation that the residents in our study are uniformly excellent leaders, and the LOFT accurately assessed this, as the group included PGY-1 residents new to clinical team leadership. Selection bias is another explanation for the frequency of high ratings, with residents who perceived themselves to be good leaders more likely to participate, and potentially selecting the team members they thought would rate them most highly. Most likely, and consistent with previous literature on feedback and assessment in medical education, raters exhibited the so-called leniency bias or generosity error out of a desire not to be negative.[21,22] This type of rater behavior may be augmented in ratings of communication and interpersonal skills, such as in a recent qualitative study of interprofessional feedback, in which health professions students attributed their hesitancy to be critical about each other's teamwork skills to discomfort with not being "nice."[23] In our study, perceived negative framing of some descriptors associated with novice leadership may have contributed to overrating, or to selecting the N/A option, which was intended to indicate that the rater had not observed a behavior. A relatively high number of raters chose the N/A option over the novice-appropriate rating, particularly when rating PGY-1 residents, and this may have inflated the ratings. Using peers or near-peer raters, rather than experts, likely aggravated this issue. Rater training has been shown to improve the reliability and validity of performance assessment ratings,[24,25] but it can be challenging when evaluators from different professional backgrounds are involved. Our evaluators had varying levels of experience in assessing the performance of residents, which may have influenced the quality of their ratings.[26] Also, that some items may not have been as easy to observe as we intended, and have been open to variable interpretations, is another explanation for the high frequency of N/A responses. We recommend that future iterations of this type of tool use a "no opportunity to observe this behavior" option instead of "N/A." Finally, we must also consider the possibility that clinical team leadership skill development does not occur along the trajectory from novice to proficient, although this would be in conflict with current thinking.[27]

Our study has a few other limitations. We only included IM and pediatrics residents on inpatient rotations, which limits the generalizability of our findings. Rater training was limited to brief instructions in the online survey. We did not collect detailed information about the thought process of raters, limiting our understanding of the response process. Finally, our study is cross-sectional, and the instrument may have performed differently if used to assess participants longitudinally.

Future work should aim to strengthen validity evidence for this instrument or an enhanced version of it, including examining whether rater training improves the performance of the instrument and how to overcome barriers to training multiprofessional raters in a clinical context. We believe the limited validity evidence of the LOFT to date does not preclude its usability for guiding formative feedback to residents on clinical team leadership. The instrument is currently being used at CUSOM to provide 360-degree feedback to residents participating in a leadership training program; preliminary data suggest the residents found it useful for this purpose (Kelsey Jones, written communication). Future research should assess whether tools that lack validity evidence may produce useful feedback that can inform performance improvement.

Conclusion

We developed, tested, and sought to provide initial validity evidence for a novel instrument to assess resident clinical leadership skills. Despite our inability to provide conclusive validity evidence for the instrument, feedback from participants and experience with its ongoing use at CUSOM suggest the

instrument has potential utility as a framework for feedback on residents' clinical leadership skills.

References

1. Accreditation Council for Graduate Medical Education. Program director guide to the common program requirements. 2012. http://www.acgme.org/Portals/0/PDFs/commonguide/CompleteGuide_v2%20.pdf. Accessed August 29, 2018.
2. Kuo AK, Thyne SM, Chen HC, et al. An innovative residency program designed to develop leaders to improve the health of children. *Acad Med*. 2010;85(10):1603–1608.
3. Kohlwes RJ, Cornett P, Dandu M, et al. Developing educators, investigators, and leaders during internal medicine residency: the area of distinction program. *J Grad Med Educ*. 2011;3(4):535–540.
4. Gurrera RJ, Dismukes R, Edwards M, et al. Preparing residents in training to become health-care leaders: a pilot project. *Acad Psychiatry*. 2014;38(6):701–705.
5. Goldman EF, Plack MM, Roche CN, et al. Learning clinical versus leadership competencies in the emergency department: strategies, challenges and supports of emergency medicine residents. *J Grad Med Educ*. 2011;3(3):320–325.
6. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *BMJ*. 2010;341:c5064.
7. Fletcher G, Flin R, McGeorge P, et al. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. *Br J Anaesth*. 2003;90(5):580–588.
8. Kim J, Neilipovitz D, Cardinal P, et al. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: the University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. *Crit Care Med*. 2006;34(8):2167–2174.
9. Clancy CM, Tornberg DN. TeamSTEPPS: assuring optimal teamwork in clinical settings. *Am J Med Qual*. 2007;22(3):214–217.
10. Frankel A, Gardner R, Maynard L, et al. Using the communication and teamwork skills (CATS) assessment to measure health care team performance. *Jt Comm J Qual Patient Saf*. 2007;33(9):549–558.
11. Malec JF, Torsher LC, Dunn WF, et al. The Mayo high performance teamwork scale: reliability and validity for evaluating key crew resource management skills. *Simul Healthc*. 2007;2(1):4–10.
12. Guise JM, Deering SH, Kanki BG, et al. Validation of a tool to measure and promote clinical teamwork. *Simul Healthc*. 2008;3(4):217–223.
13. Kim J, Neilipovitz D, Cardinal P, et al. 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.
14. Grant EC, Grant VJ, Bhanji F, et al. The development and assessment of an evaluation tool for pediatric resident competence in leading simulated pediatric resuscitations. *Resuscitation*. 2012;83(7):887–893.
15. Posner BZ, Kouzes JM. Development and validation of the leadership practices inventory. *Educ Psychol Meas*. 1988;48(2):483–496.
16. Messick S. The psychology of educational measurement. *J Educ Meas*. 1984;21(3):215–237.
17. American Educational Research Association; American Psychological Association; National Council on Measurement in Education; Joint Committee on Standards for Educational and Psychological Testing. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 2014.
18. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. 2006;3(2):77–101.
19. Waggoner J, Carline JD, Durning SJ. Is there a consensus on consensus methodology? Descriptions and recommendations for future consensus research. *Acad Med*. 2016;91(5):663–668.
20. Penfield R, Miller J. Improving content validation studies using an asymmetric confidence interval for the mean of expert ratings. *Appl Meas Educ*. 2004;17:359–370.
21. Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. *Acad Med*. 2005;80(10):84–87.
22. Bandiera G, Lendrum D. Daily encounter cards facilitate competency-based feedback while leniency bias persists. *CJEM*. 2008;10(1):44–50.
23. Mandal J, Avdagic K, Wamsley M, et al. Beyond "great job": feedback among students on interprofessional teams. *J Interprof Educ Pract*. 2016;5:37–43.
24. Woehr DJ, Huffcutt AI. Rater training for performance appraisal: a quantitative review. *J Occup Organ Psychol*. 1994;67:189–205.
25. Evans LV, Morse JL, Hamann CJ, et al. The development of an independent rater system to assess residents' competence in invasive procedures. *Acad Med*. 2009;84(8):1135–1143.
26. Govaerts MJB, Schuwirth LWT, Van der Vleuten CPM, et al. Workplace-based assessment: effects of rater expertise. *Adv Health Sci Educ Theory Pract*. 2011;16(2):151–165.
27. Lord RG, Hall RJ. Identity, deep structure and the development of leadership skill. *Leadership Q*. 2005;16:591–615.

Sandra K. Oza, MD, MA, is Assistant Professor of Medicine, Department of Medicine, Albert Einstein College of Medicine and

Montefiore Medical Center; **Sandrijn van Schaik, MD, PhD,** is Associate Professor of Pediatrics, Department of Pediatrics, University of California, San Francisco School of Medicine; **Christy K. Boscardin, PhD,** is Associate Professor of Medicine, Department of Medicine, University of California, San Francisco School of Medicine; **Read Pierce, MD,** is Assistant Professor of Medicine, Department of Medicine, University of Colorado School of Medicine; **Edna Miao, MD,** is a Resident Physician, Department of Pediatrics, Loma Linda Children's Hospital; **Tai Lockspeiser, MD, MHPE,** is Associate Professor of Pediatrics, Department of Pediatrics, University of Colorado School of Medicine; **Darlene Tad-y, MD,** is Associate Professor of Medicine, Department of Medicine, University of Colorado School of Medicine; **Eva Aagaard, MD,** is Professor of Medicine and Senior Associate Dean for Education, Department of Medicine, Washington University School of Medicine; and **Anda K. Kuo, MD,** is Professor of Pediatrics, Department of Pediatrics, University of California, San Francisco School of Medicine.

Funding: This work was funded by a grant from the University of California, San Francisco Academy of Medical Educators.

Conflict of interest: The authors declare they have no competing interests.

The authors would like to thank the expert panelists who contributed to phase 2 of this study, as well as the resident participants and team evaluators who participated in all phases of the study.

Portions of this work have previously been presented in abstract form at the following meetings: the UCSF Education Symposium, San Francisco, California, April 2013; the Society of General Internal Medicine Annual Meeting, Denver, Colorado, April 25–27, 2013; the Association of American Medical Colleges Western Group on Educational Affairs Regional Meeting, Irvine, California, May 4–7, 2013; and the Association of American Medical Colleges Annual Meeting, Philadelphia, Pennsylvania, November 1–6, 2013.

Corresponding author: Sandra K. Oza, MD, MA, R. L. Gottesman Clinical Skills Center, Van Etten Building 2B-06, 1300 Morris Park Avenue, Bronx, NY 10461, 718.862.1770, koza@montefiore.org

Received February 1, 2018; revisions received May 17, 2018, and July 6, 2018; accepted July 17, 2018.