The Mindful Manager: Validation of a Rounding Leadership Instrument for Residents

Daniel N. Ricotta, MD^{1,2,3} , Brittany L. Ranchoff, MPH^{1,3}, Christine P. Beltran^{1,3}, Andrew J. Hale, MD⁴, Jason A. Freed, MD^{1,2}, and Grace C. Huang, MD^{1,2,3}

¹Department of Medicine Beth, Israel Deaconess Medical Center, Boston, MA, USA; ²Harvard Medical School, Boston, MA, USA; ³Carl J. Shapiro Institute for Education and Research, Boston, MA, USA; ⁴Larner College of Medicine at the University of Vermont, Burlington, VT, USA.

BACKGROUND: In the context of inpatient general medicine, “rounding” refers to the process of seeing, assessing, and caring for patients as a team. The clinical leadership skills required of residents to lead rounds are essential to inpatient care and clinical education. Assessment of these skills has relevance to developing competent physicians; however, there is an absence of widely accepted tools to specifically measure this competency.

OBJECTIVE: To develop and collect validity evidence for a direct observation instrument of internal medicine residents’ leadership skills during daily inpatient care rounds for future formative assessment.

DESIGN: Prospective observational study.

PARTICIPANTS: PGY2 and PGY3 internal medicine residents.

MAIN MEASURES: The authors collected inferences of validity evidence according to Kane’s validity model. They performed direct observations of PGY2 and PGY3 residents by individual faculty and trained raters and measured inter-rater reliability, using the kappa statistic. Mixed linear regression models were used to compare PGY2 and PGY3 residents. Surveys captured faculty perceptions about value of the instrument.

KEY RESULTS: A total of 223 observations were performed in 92 unique individuals. Twenty-four faculty used the observation instrument, of which 18 (75%) completed the post-survey, and 100% agreed that the instrument represented the resident’s global leadership abilities. Inter-rater reliability was strong, with an overall kappa statistic equaling 0.82. The mean performance for PGY2 and PGY3 residents was 15.9 (SD 5.1) and 17.7 (SD 4.1), respectively. Adjusting for repeated measures, there was no statistically significant difference between groups.

CONCLUSIONS: The authors reported evidence for all four stages of validity and use of the instrument in clinical practice. Their work provides a codification of best practices of rounding leadership, which directly impacts the education of trainees, care of hospitalized patients, and use for formative assessment. The instrument also has the potential to be used for summative assessment.

KEY WORDS: validation; assessment; graduate medical education; leadership.

J Gen Intern Med 35(4):1161–6

DOI: 10.1007/s11606-019-05348-1

© Society of General Internal Medicine 2019

INTRODUCTION

Every day across the USA, teams of internal medicine (IM) physicians, trainees, and other healthcare professionals are assessing and caring for hospitalized medical patients. “Rounds,” as this inpatient care model has come to be known, represents the foundation of inpatient care and clinical education.^{1, 2} Inpatient teams comprising an attending physician, resident, interns, and medical students are generally led by the resident in most academic centers. The resident leader determines which patients and in what order the team sees together, and is responsible for bedside teaching, managing intra-team conflict, and facilitating communication.

Given the central role residents play in inpatient care delivery and clinical education, rounding leadership skills are essential to post-graduate medical training.^{3, 4} IM graduate training programs are expected to educate resident physicians to practice independently and lead patient care teams.⁵ At most training programs, however, clinical leadership skills vital to inpatient care such as task delegation, time management, and clinical debriefing are acquired predominantly through an apprenticeship model.^{6–8} Skill acquisition through this means results in a variable set of practices which may impair both learner experiences and patient outcomes.

One factor contributing to the challenge of developing and evaluating leadership curricula is the absence of widely accepted tools that specifically assess the quality of how residents run rounds.⁹ According to a recent systematic review, the majority of leadership instruments assess the function of a team as a whole and focus on discrete events such as a medical code or trauma.⁹ These types of assessments are often time-limited, examine singular outcomes such as return of spontaneous circulation, and can be too inflexible or algorithmic to meet the heterogeneity and day-to-day variability of daily rounds. This is important because existing objective clinical leadership assessments do not reflect the primary setting of residency training: routine daily clinical care. To our knowledge, no published instrument has been developed for the sole purpose of assessing

Received April 4, 2019

Accepted September 10, 2019

Published online January 2, 2020

individual resident leadership during daily rounds, namely, rounding leadership. Approaches attempting to adapt current leadership assessments to daily rounds have been limited by variability, consistency, and validity.¹⁰ Further, alternative methods often rely on subjective measures and are prone to bias and halo effect,^{11, 12} while others have not been used in the clinical setting to capture real-life performance. Finally, there is an absence of codified best practices¹³ in rounding leadership to serve as guiding frameworks.

Therefore, the purpose of our study was to develop an instrument to measure resident rounding leadership skills for formative feedback and to collect validity evidence based on Kane's 4-stage model.¹⁴

METHODS

We conducted this study at Beth Israel Deaconess Medical Center (BIDMC), a 673-bed quaternary-care academic medical center in Boston, MA, between March 2016 and August 2018. The BIDMC IM residency program included 165 total residents, of which there were 52 post-graduate year 2 (PGY2) and 48 post-graduate year 3 (PGY3) residents. The inpatient general medicine teams comprised a single PGY2 or PGY3 resident, a single hospital medicine attending, two interns, and one or two medical students. The sample size was determined by the number of residents assigned to a general medicine service during the study period.

Instrument Development

We first performed a literature review by searching PubMed, Google Scholar, and non-peer-reviewed literature for medical and non-medical frameworks of effective team leadership skills.^{5, 15–27} Two authors (DR, GH) with expertise in both inpatient medicine and residency training extrapolated only the aspects that were directly applicable to managing inpatient medical teams during daily rounds, such as setting daily expectations, delegating tasks explicitly, and role-modeling behaviors.

We then categorized the leadership skills into broad domains, which were subsequently reviewed by residency program leadership to further narrow and define domains relevant to resident practice, important to residency training, and aligned with expectations of the Accreditation Council of Graduate Medical Education (ACGME).¹³

We then sought consensus from a group of eight BIDMC hospital medicine core faculty, defined as hospitalists with at least 5 years of experience in an academic medical center, and practicing primarily on teaching services. Following three rounds of review and revision, we achieved consensus on six domains of rounding leadership skills: (1) creating a safe learning environment; (2) providing consistent and effective feedback; (3) role-modeling clinical skills, attitudes, and behaviors; (4) running the team efficiently; (5) developing situational awareness; and (6) teaching.

Next, we created behavioral elements for each of the six domains that could be observed objectively on a daily basis, based on anecdotal experience. The preliminary instrument was then circulated among core faculty and program leadership for two additional rounds of input until consensus was achieved.

We then piloted the preliminary instrument among five hospitalists for 10 different residents over a 4-month period (phase 1, instrument development and testing). We consequently refined the instrument further, and we added a subjective scale element for "exemplary" at the request of hospitalists in order to stratify residents further. Residency program leadership approved the final instrument which we then built into Qualtrics® survey software (Provo, UT) for ease of use.

Analysis

We employed Kane's four-stage model to organize the collection of validity evidence for our instrument in accordance with its increasing use in medical education and focus on process of development.²⁸ The framework entails 4 stages of inference: scoring, how the observation results in a reproducible and quantifiable result; generalization, a measure of reliability and how the sample items reflect the performance as a whole; extrapolation, the correlation between testing and real-world performance; and implication, the interpretation of, and subsequent action about, the results.^{14, 28}

As described by Kane, prior to examining the four stages of inference, we defined the purpose of the instrument, namely to objectively measure rounding leadership skills for formative assessment and coaching of the resident. Additionally, we recognized that a total score could allow residents to compare themselves against an objective metric and to monitor individual progress over time, and the results could also be used to identify systemic performance gaps across housestaff in order to inform future leadership curricula. We explored the stages of validity evidence in three phases: Phase 1 included instrument development, pilot testing, and rater training between March 2016 and June 2017, phase 2 comprised faculty-only observations of PGY2–3 residents between July and August of 2017, phase 3 introduced paired observations of PGY2–3 residents by both a research assistant (RA) and faculty member during February–March and July–August of 2018. A summary of the steps for instrument development and testing is shown in Figure 1.

Scoring

Scoring, as described by Kane, refers to the translation of an observation into a score that is both an accurate representation of the individual's performance and is reproducible. The elements of the scoring rubric were determined through literature review, initial development, consensus building, and pilot testing, as described above. We then assessed this inference in phase 2 of instrument development by employing faculty raters to score resident performance using the instrument. Faculty raters were of two types: supervising hospitalists on clinical service and faculty who routinely attended rounds but

Figure 1 Flowchart representing the phases of instrument development and process of gathering validity evidence. RA research assistants.

were not part of patient care, such as assigned teaching attendings. All faculty raters underwent rater training in the forms of an overview at a hospitalist staff meeting, a more in-depth faculty development seminar where the logistics of observing the residents were presented, and an in-person, individual review of the instrument. During the individual review, each of the instrument items was reviewed and examples given of the presence and absence of each item. In addition, all raters received periodic email reminders, including references on how to use the instrument. A subset of faculty raters received additional face-to-face training, to clarify any confusion and score additional hypothetical scenarios, at their request.

Faculty raters completed the instrument following a single morning of rounds. Each item on the instrument was worth 1 point scored by a binary outcome of “observed” or “not observed” for a total possible score of 26. After completing observations, the faculty raters completed a perception survey that solicited their feedback about the instrument. The survey addressed ease of use and relevance to clinical practice. Raters were allowed to provide feedback to the resident following the encounter, after the instrument was completed and submitted.

Generalization

Kane describes generalization as performance in a testing setting. He suggests that evidence of generalization comes from the process of ensuring that chosen items reflect an appropriate sample and that scores are reliable. The intent of

our instrument was for use in a real-world setting with expected day-to-day variability in experience; therefore, we focused on the consensus-driven process of choosing behavioral elements and inter-rater reliability to establish evidence of generalization. To this end, we employed faculty raters and RAs to observe residents in pairs during phase 3. Each independently completed the instrument following a rounding experience, blinded to the others’ scoring. Items were dichotomized as observed or not observed. Results were then analyzed using the kappa statistic to assess inter-rater reliability by item.

Extrapolation

Extrapolation refers to the relationship between scores and real-world performance. Evidence to support extrapolation comes from the methods used to ensure the instrument reflects essential components of real-world performance and empiric testing to assess this relationship.

To measure correlation with real-world performance, we compared performance of PGY2 and PGY3 residents from phase 2 observations, based on the hypothesis that more experience would result in higher scores. Additionally, faculty raters completed a survey to assess whether they felt the instrument reflected the skills of rounding leadership.

Statistical Analysis

We used the chi-square and Fisher’s exact tests to analyze proportional data. We used the kappa statistic to assess the inter-rater reliability, where strong reliability was defined as kappa > 0.8.²⁹ We used mixed linear regression models to compare PGY2 and PGY3 residents, using a random effect for learner identifier to control for multiple evaluations per learner. Analyses were performed using JMP® Pro, Version 14 (SAS Institute Inc., Cary, NC).

This study was approved by the BIDMC Institutional Review Board as exempt educational research.

RESULTS

Scoring

Between phases 2 and 3, there were 223 observations of 92 unique individuals. Of these observations, 140 were of 63 PGY2 residents, while the remaining were of PGY3 residents. Twenty-four different faculty completed the observation instrument, of which 18 (75%) completed the post-survey. Results from faculty surveys were overwhelmingly positive as described in Table 1. The most common selected response to the duration of time to complete the instrument was “1–3 minutes.”

Generalization

During phase 3, there were 36 paired observations of 23 unique individuals (12 of whom were PGY2). Inter-rater reliability for each item is reported in Table 2, and the average

Table 1 Results from 18 (75%) Faculty Who Completed a Perception Survey About the Instrument at Beth Israel Deaconess Medical Center in 2018

Question	Agree or strongly agree, n (%)
The instrument was easy to complete	17 (94%)
The behavioral elements were objective	16 (89%)
The behavioral elements were easy to observe	16 (89%)
The behavioral elements represent best practices of managing an inpatient medical team	17 (94%)
The instrument was helpful in providing feedback to the resident	17 (94%)
The instrument was helpful in assessing the resident	18 (100%)
The instrument is clinically relevant to inpatient care teams	18 (100%)
The instrument provides an accurate representation of the resident's global leadership abilities	18 (100%)

kappa for the entire instrument was 0.82. Of the 26 items, 4 items had low kappa scores: Q1, Q2, Q13, and Q26.

Extrapolation

We designed our instrument for observation of real-world encounters, and therefore correlation of test environments to actual performance was not applicable. Instead, evidence of extrapolation for our study came from the approach we used to develop the instrument, namely, the evidence base from the literature, iterations of expert review with consensus, and pilot-testing, and our findings that faculty agreed with interpretations of the instrument.

In phase 2, there were a total of 49 PGY2 and 23 PGY3 observations. The mean scores were 15.9 (SD 5.1) and 17.7 (SD 4.1) for PGY2 and PGY3 respectively. Adjusting for repeated measures, the results of the random mixed model revealed no statistically significant difference between groups ($p = 0.22$).

Additionally, data from our faculty survey (Table 1) provided evidence that the instrument “reflects the key aspects of performance” and “stakeholders agree with the interpretations and anticipate that they will apply the new contexts in training or practice.”²⁸ Specifically, 94% of faculty reported that the instrument elements represented best practices of managing an inpatient medical team, 100% reported the instrument reflected the residents’ global leadership skills, and 83% applied their interpretations of the residents’ behaviors to coaching.

Implications

The last stage of Kane’s model considers downstream consequences that arise from the instrument’s scores. Kane suggests that the proposed action should inform the degree of scrutiny necessary to support that intention. For example, an instrument used for formative assessment will not require as much examination of the consequences as a high-stakes

Table 2 Inter-rater Reliability by Instrument Item Categorized by Domains for 36 Paired Observations Performed at Beth Israel Deaconess Medical Center in 2018

Instrument item	Inter-rater reliability (kappa)
Q1: Comes to rounds already with a sense of what clinical plan should be for each patient	0.44
Creates a safe learning environment	
Q2: Sets expectations—at the start of rotation or on a daily basis	0.39
Q3: Asks questions in a non-threatening way	1.0
Q4: Provides support and encouragement	0.84
Q5: Incorporates all members of the healthcare team	1.0
Q6: Identifies learners needs and knowledge gaps	0.77
Maintains situational awareness	
Q7: Sets agenda—verbalizes a rounding order	0.91
Q8: Sets agenda—identifies strategies for rounding	0.87
Q9: Sets agenda—risk stratifies patients for rounding (dire, dilemma, discharge, delineated)	0.94
Q10: Acknowledges needs of the team	0.94
Q11: Limits interruptions	0.94
Role models	
Q12: Explicitly verbalizes own specific behaviors or actions	1.0
Q13: Demonstrates effective bedside manner	0.47
Q14: Avoids pejorative comments in front of learners	0.77
Efficiency	
Q15: Starts rounds on time	1.0
Q16: Ends rounds at a specified time point	0.89
Q17: Redirects learners to avoid tangential or circumferential discussion	0.77
Q18: Summarizes plan at the end of clinical encounters	0.94
Q19: Explicitly assigns tasks, delegates responsibilities	1.0
Teaching	
Q20: Uses open ended “why” and “how” questions as opposed to “what” questions	0.89
Q21: Uses body language effectively to promote learner engagement	0.72
Q22: Teaching is incorporated into clinical care	1.0
Feedback	
Q23: Provides immediate feedback during or immediately after rounds	0.83
Q24: Feedback is specific and actionable	0.92
Q25: Debriefs after clinical encounters	0.94
Q26: Checks in with team daily	0.18

examination, such as board-certification. Our instrument was only used for feedback and coaching and 15 (83%) faculty reported using the instrument to inform coaching of the resident. Furthermore, the results of the observations were stored securely, none of the observations were provided to the resident program director, reported for ACGME milestones, or discussed during competency meetings. Therefore, the fact that results did not affect professional standing addressed this validity inference.

DISCUSSION

We developed an instrument to measure resident rounding leadership skills. In doing so, we identified the relevant constructs of clinical leadership for daily rounds—creating a safe learning environment, regular team debriefings, role-

modeling, efficiency, developing situational awareness, and teaching—based on a review of the literature, synthesized through a consensus-driven process, and refined through iterations of pilot testing. We trained hospitalists and research assistants as standardized raters and demonstrated strong inter-rater reliability. We found evidence in all four stages of Kane's validity inference and for use of the instrument in clinical practice. Faculty raters felt the instrument reflected the overall leadership skills of the individual, was relevant to clinical practice, was easy to use, and took minimal time to complete. Moreover, nearly all faculty used the instrument for feedback and coaching following observations.

We found strong inter-rater reliability for most items on the instrument which likely reflected the in-depth and iterative process of construct development. Additionally, the high level of reliability may speak to the importance of a multi-faceted approach to faculty training, which we achieved through small- and large-group sessions, along with voluntary calibration training. Our findings of four items with poor inter-rater reliability (comes to rounds prepared, demonstrates good bedside manner, sets expectations, checks-in with teams daily) were possibly explained by the inherent differences between research assistants and faculty raters. For instance, faculty are more likely to recognize when residents come prepared to rounds due to tacit knowledge. Research assistants were often not permitted to enter the room of patients as they were not involved in patient care, and so would have lost the opportunity to observe bedside manner. Lastly, work rounds occasionally began informally before research assistants arrived, and daily check-ins sometimes occurred outside of morning rounds, where the observations exclusively occurred.

While we compared novice and experienced learners as evidence for extrapolation validity, it is not surprising that we did not detect a significant difference in performance. We suspect that these results arose from the fact that our training program did not yet have a formal curriculum for rounding leadership. Previously, we relied on an apprenticeship model to teach leadership skills, which may be unstandardized and of variable quality and thus unlikely to mature with experience alone. It is also possible that rounding leadership skills do not progress in the same fashion as other clinical skills.³⁰

Our work builds on prior literature assessing leadership in the clinical arena, while specifying it in the context of daily rounds. Related instruments include the Leadership Observation and Feedback Tool (LOFT), an adaptation of the Leadership Practice Inventory from business literature to the clinical environment.³⁰ The 10 domains described in the LOFT had similarities to our work, such as creating a safe learning environment, providing constructive feedback, role-modeling, and setting expectations. Unlike our instrument, the authors of the LOFT instrument reported a strong halo effect and concluded there was insufficient validity evidence to support its use for measurement, though, similar to our findings, they did find it useful for providing feedback.³⁰ Another group applied 9 different teamwork observation instruments designed for

other settings to the context of daily rounds and found discordant assessments across instruments.¹⁰ They concluded that existing frameworks may not be suitable for daily rounds and recommended the development of novel instruments for rounding leadership, and collecting validity evidence, which we have accomplished here.

Limitations to our work include the single institutional setting and potential lack of generalizability to other programs such as community-based and non-IM residency programs. We did not formally use a Delphi method for consensus-building of the leadership domains. Behaviors observed on daily rounds also may vary day-to-day due to factors inherent in complex healthcare environments, such as clinical volume and patient acuity. While residents were not told when the instrument was being used, it is possible that residents altered their behavior when a rater was present. Application to other settings will require robust faculty development, which may not exist in other institutions. Without rater training, the instrument may be prone to halo effect, as documented in other assessment methods. We did not formally measure the impact of using the instrument on skill development as part of the implications phase, and our lack of statistical significance comparing PGY2 and PGY3 residents may signal the insensitivity of our instrument to detect change. Practical aspects of our study also limited the sample size available to us and may have resulted in the lack of power to detect small differences. We did not directly compare our instrument to others to assess correlation. Finally, we did not end up including an "exemplary" rating in our analysis, due to infrequency of its use, which may limit its ability to discriminate between resident skill and the quality or frequency of behavioral elements observed.

CONCLUSIONS

Rounding leadership is foundational to medical training, is challenging to measure, and can directly influence the clinical learning environment. We aimed to codify best practices of clinical leadership in the context of daily inpatient rounds founded on Kane's validity model. While we used the instrument for feedback, there may be opportunity for use as summative milestones-based assessment and further examination of its correlation with patient outcomes. Other potential applications include use for subspecialty services and for faculty development in leadership. On a psychometric level, adding global rating scales to the instrument would offer additional variables with which to collect validity evidence. Our results will also form the basis of evaluating a resident-as-leader curriculum for rising PGY2 residents. Leadership has emerged as a core competency of physicians in the current era of team-based care, and objective assessment of these skills may improve how we train internal medicine residents to care for patients in our complex healthcare system.

Acknowledgments: The authors greatly appreciate the work of Amy Sullivan EdD and Amy Cohen EdM for their assistance in data analysis.

Corresponding Author: Daniel N. Ricotta, MD; Department of Medicine Beth Israel Deaconess Medical Center, Boston, MA, USA (e-mail: dricotta@bidmc.harvard.edu).

Compliance with Ethical Standards:

Ethical Approval: The BIDMC IRB approved this study as exempt educational research.

Conflict of Interest: The authors declare that they do not have a conflict of interest.

REFERENCES

1. Calderon AS, Blackmore CC, Williams BL, et al. Transforming Ward Rounds Through Rounding-in-Flow. *J Grad Med Educ*. 2014;6(4):750–755. doi:<https://doi.org/10.4300/JGME-D-13-00324.1>.
2. Zabar S, Hanley K, Stevens DL, et al. Measuring the competence of residents as teachers. *J Gen Intern Med*. 2004;19(5 Pt 2):530–533. doi:<https://doi.org/10.1111/j.1525-1497.2004.30219.x>.
3. Bing-You RG, Sproul MS. Medical students' perceptions of themselves and residents as teachers. *Med Teach*. 1992;14(2–3):133–138.
4. Smith CC, McCormick I, Huang GC. The Clinician-Educator Track. *Acad Med*. 2014;89(6):888–891. doi:<https://doi.org/10.1097/ACM.0000000000000242>.
5. ACGME & ABIM. The Internal Medicine Milestone Project. *Intern Med Milestone Proj*. 2013;(July):1–28.
6. Baird DS, Soldanska M, Anderson B, Miller JJ, York N. Current leadership training in dermatology residency programs: A survey. *J Am Acad Dermatol*. 2012;66(4):622–625. doi:<https://doi.org/10.1016/j.jaad.2009.10.009>.
7. Blumenthal DM, Bernard K, Bohnen J, Bohmer R. Addressing the Leadership Gap in Medicine. *Acad Med*. 2012;87(4):513–522. doi:<https://doi.org/10.1097/ACM.0b013e31824a0c47>.
8. Adams J. The elusive leadership competency. *J Grad Med Educ*. 2011;3(3):442–443. doi:<https://doi.org/10.4300/JGME-D-11-00162.1>.
9. Rosenman ED, Ilgen JS, Shandro JR, Harper AL, Fernandez R. A systematic review of tools used to assess team leadership in health care action teams. *Acad Med*. 2015;XX(X):1–15. doi:<https://doi.org/10.1097/ACM.0000000000000848>.
10. Weingart SN, Yaghi O, Wetherell M, Sweeney M. Measuring Medical Housestaff Teamwork Performance Using Multiple Direct Observation Instruments. *Acad Med*. 2018;93(7):1064–1070. doi:<https://doi.org/10.1097/ACM.0000000000002238>.
11. Gray JD. Global rating scales in residency education. *Acad Med*. 1996;71(1 Suppl):S55–63.
12. Thorndike EL. A constant error in psychological ratings. *J Appl Psychol*. 1920;4(1):25–29. doi:<https://doi.org/10.1037/h0071663>.
13. Jardine D, Correa R, Schultz H, et al. The Need for a Leadership Curriculum for Residents. *J Grad Med Educ*. 2015;(ACGME News and Views):307–309. doi:<https://doi.org/10.4300/JGME-07-02-31>.
14. Kane M. Validation. In: Praeger, ed. *Educational Measurement*. 4th ed. Westport, CT: Brennan RL, ed; 2006:17–64.
15. Bannister SL, Wickenheiser HM, Keegan DA. Key Elements of Highly Effective Teams. *Pediatrics*. 2014;133(2):184–186. doi:<https://doi.org/10.1542/peds.2013-3734>.
16. Agency for Healthcare Research and Quality A. TeamSTEPPS 2.0. *Pocket Guid*. 2013:1–36.
17. Carbo AR, Tess A V, Roy C, Weingart SN. Developing a High-Performance Team Training Framework for Internal Medicine Residents. *J Patient Saf*. 2011;7(2):72–76. doi:<https://doi.org/10.1097/PTS.0b013e31820dbe02>.
18. Gabel S. Expanding the scope of leadership training in medicine. *Acad Med*. 2014;89(6):848–852. doi:<https://doi.org/10.1097/ACM.0000000000000236>.
19. Taylor CA, Taylor JC, Stoller JK. Exploring Leadership Competencies in Established and Aspiring Physician Leaders: An Interview-based Study. *J Gen Intern Med*. 2008;23(6):748–754. doi:<https://doi.org/10.1007/s11606-008-0565-5>.
20. Lee TH. Turning doctors into leaders. *Harv Bus Rev*. 2010;88(4):50–58.
21. RW G. *Fundamentals of Management*. 5th ed. Boston: Mass Houghton Mifflin Company; 2008.
22. Zak PJ. The neuroscience of Trust: Management Behaviors That Foster Employee Engagement. *Harv Bus Rev*. 2017;(January-February):86–90.
23. Lytkina BE, Powell KR, Kincaid S, Wang D. What Sets Successful CEOs Apart: The Four Essential Behaviors That Help Them Win the Top Job and Thrive Once They Get It. *Harv Bus Rev*. 2017;95(3):70–77.
24. Dine CJ, Kahn JM, Abella BS, Asch DA, Shea JA. Key elements of clinical physician leadership at an academic medical center. *J Grad Med Educ*. 2011;3(1):31–36. doi:<https://doi.org/10.4300/JGME-D-10-00017.1>.
25. Stoller JK, Rose M, Lee R, Dolgan C, Hoogwerf BJ. Teambuilding and Leadership Training in an Internal Medicine Residency Training Program Experience with a One-day Retreat. *JGIM*. 2004;19:692–697.
26. Shrader PharmD S, Kern MD D, Zoller PhD J, Blue PhD A. Interprofessional Teamwork Skills as Predictors of Clinical Outcomes in a Simulated Healthcare Setting. *J Allied Health*. 2013;42(1):e1–6.
27. Kiesewetter J, Schmidt-Huber M, Netzel J, Krohn AC, Angstwurm M, Fischer MR. Training of leadership skills in medical education. *GMS Z Med Ausbild*. 2013;30(4):Doc49.
28. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: A practical guide to Kane's framework. *Med Educ*. 2015;49(6):560–575. doi:<https://doi.org/10.1111/medu.12678>.
29. McHugh ML. Interrater reliability: the kappa statistic. *Biochem medica*. 2012;22(3):276–282.
30. Oza SK, Van schaik S, Boscardin CK, et al. Leadership Observation and Feedback Tool : A Novel Instrument for Assessment of Clinical Leadership Skills. *J Grad Med Educ*. 2018;10(5):573–582. doi:<https://doi.org/10.4300/JGME-D-18-00113.1>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.