Development and Empirical Testing of a Novel Team Leadership Assessment Measure: A Pilot Study Using Simulated and Live Patient Encounters

Elizabeth D. Rosenman, MD, Mark J. Bullard, MD, MS-HPed, Kerin A. Jones, MD, MS, Laura Welsh, MD, Sarah M. Brolliar, BS, Benjamin R. Levine, MS, James A. Grand, PhD, and Rosemarie Fernandez, MD

ABSTRACT

Background: Team leadership is critical to health care resuscitation team performance. There has been increased focus on competency in team leadership behaviors; however, there is still variability in how team leadership is assessed within emergency medicine. The objective of this study was to develop and pilot a novel team leadership assessment measure for emergency medicine resuscitation teams.

Methods: Team leadership dimensions and associated behaviors were identified through a systematic literature review and expert consensus. Included behaviors were used to create behaviorally anchored rating scales, which were then revised based on subject matter expert ratings. Four raters from three different academic institutions observed 30 video-recorded resuscitations (20 simulated and 10 actual patient care resuscitations). Mean leadership scores were calculated. Intraclass coefficients (ICCs) were calculated for each item and for overall leadership scores. Leader scores for the simulation-based scenarios were compared to external variables including level of training, team process, clinical performance, and team situational awareness. The study was conducted from July 2017 through June 2018.

From the Department of Emergency Medicine, University of Washington School of Medicine (EDR, SMB, LW, RF), Seattle, WA; the Department of Emergency Medicine, Carolinas Medical Center, Atrium Health (MJB), Charlotte, NC; the Department of Emergency Medicine, Wayne State University (KAJ), Detroit, MI; the Department of Psychology, University of Maryland (BRL, JAG), College Park, MD. Dr. Welsh is currently with the Department of Emergency Medicine, Boston University School of Medicine, Boston, MA. Dr. Fernandez is currently with the Department of Emergency Medicine, University of Florida College of Medicine, Jacksonville, FL.

Received October 5, 2018; revision received December 20, 2018; accepted January 2, 2019.

Funding and support for this project was provided by the Society for Academic Emergency Medicine Foundation (EF2017-001, [EDR]) and the Agency for Healthcare Research and Quality (1R18HS020295 [RF, EDR] and 1R18HS022458-01A1 [RF, EDR]). The funding source had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; or preparation, approval, or decision to submit the manuscript.

EDR reports grant money to the University of Washington to conduct research conceived and written by RF from the Agency for Healthcare Research and Quality and the Department of Defense. EDR has received funding personally from Physio-Control, Inc., for speaker fees. LW reports grant money to the University of Washington to conduct research conceived and written by LW from the Association of American Medical Colleges' Western Group on Educational Affairs. SMB reports grant money to the University of Washington to conduct research conceived and written by RF from the Agency for Healthcare Research and Quality and the Department of Defense. JAG reports grant money to the University of Maryland to conduct research conceived and written by JAG and RF from the Department of Defense. RF reports grant money to the University of Washington and the University of Florida to conduct research conceived and written by RF from the Agency for Healthcare Research and Quality, the Department of Defense, and the State of Washington Department of Labor and Industry. RF has received funding personally from Physio-Control, Inc., for speaker fees. The other authors have disclosed no conflicts of interest.

Author contributions: study concept and design—EDR and RF; acquisition of data—EDR, MJB, KAJ, LW, and SMB; analysis and interpretation of the data—EDR, SMB, BRL, and JAG; drafting of the manuscript—EDR; critical revision of the manuscript—MJB, KAJ, LW, and RF; statistical expertise—BRL and JAG; and acquisition of funding—EDR and RF.

Supervising Editor: Daniel Egan, MD.

Address for correspondence and reprints: Elizabeth D. Rosenman, MD; e-mail: er24@uw.edu.

AEM EDUCATION AND TRAINING 2019;3:163–171

Results: Leadership scores ranged from 2.23 to 4.30 (mean [$\pm$ SD] = 3.18 [$\pm$ 0.50]). The ICC for the overall score was 0.79 for all observations, 0.87 for simulation-based observations, and 0.24 for the patient care observations. Team leadership scores on simulation-based observations did not correlate with available external variables.

Conclusions: We developed a novel team leadership assessment measure for emergency medicine resuscitation teams with supporting validity evidence, including content validity and response process. The measure demonstrated acceptable inter-rater reliability when applied to simulation-based medical resuscitations; however, this did not translate to trauma resuscitations in the actual patient care setting.

Effective team leadership is linked to better teamwork,¹ and better teamwork is thought to improve patient care.² Leadership is especially critical in health care action teams, such as emergency medicine (EM) resuscitation teams, that function under complex, dynamic, and time-pressured conditions.³ In action teams, leadership is conceptualized as functional, defined as facilitating and coordinating individual and team efforts to accomplish shared goals.⁴ Isolating team leader performance is complicated, as the performance of any individual within a team is influenced by interactions with the team itself.^{4,5} Despite the interdependence of teams, educators are expected to assess and develop individual performance during residency training. The importance of training and assessing team leaders has been emphasized within graduate medical education as a whole^{6,7} and within EM specifically.^{8,9}

Developing competent team leaders requires robust assessment tools; however, there is variability in how leadership skills are assessed throughout medicine and very little work specific to EM.¹⁰ The majority of published assessment measures include leadership as one component of teamwork, with very few measures focused specifically on leadership. The most commonly published leadership-focused measure, the adapted leadership behavior description questionnaire (LBDQ),¹¹ has been modified with almost every published application of the measure, preventing an aggregate understanding of its utility and evidence of validity. For example, it has been applied as a checklist, as ratings scales with frequency anchors, and as the frequency of related leadership utterances.¹⁰ Several other leadership-focused measures also utilize rating scales with minimal or no anchors, making the measure dependent on rater expertise and consistency.^{12,13} Other limitations of existing leadership assessment measures include application to a narrow clinical focus (e.g., trauma care),^{12,13} empirical testing and validation only on simulation-based care,¹⁴ and assessing leadership at the team level rather than focusing on the performance of an individual team leader.¹⁵

Simulation is frequently recommended as a platform for assessment in graduate medical education.¹⁶ Advantages include the ability to standardize the assessment across learners and the relative ease of using dedicated observers or video recording. Transfer of performance from simulation to the bedside, however, is not guaranteed and EM educators still need the ability to accurately and reliably measure team leadership during actual patient care. The objective of this pilot study was to develop and empirically test a conceptually sound EM resuscitation team leader assessment measure that could be used in both simulated and actual patient care settings.

METHODS

Overview

We used a three-phase process to create and evaluate a novel team leader assessment measure: 1) initial measure development, 2) measure review and refinement, and 3) empirical testing of the measure (Figure 1).^{17,18} We collected validity evidence from four sources (content validity, internal structure, response process, and relationship to other variables)^{19,20} by applying the measure to medical and trauma resuscitations in simulated and actual patient care events. The University of Washington Institutional Review Board approved this study. The study occurred over a 12-month period, from July 2017 through June 2018.

Phase 1: Initial Measure Development

We created a list of leadership behaviors from two existing systematic reviews of health care team leadership.^{10,21} We supplemented this list using studies published since these reviews (2012–2017)^{13,14,22–30} and the EM Milestone Project.¹⁶ The behaviors were then organized into team leadership dimensions. An existing taxonomy of teamwork behaviors served as a conceptual framework for this process.³¹ A group of six subject matter experts (SMEs) independently reviewed the dimensions and the leadership behaviors for relevance and importance to EM team leadership using

Figure 1. Overview of the three-phase process used for measure design and testing. BARS = behaviorally anchored rating scales; SME = subject matter expert.

the Delphi method.³² A sample size of six SMEs was selected based on previously reported standards for content validation.³³ The SMEs represented five different institutions and included two organizational psychologists with an expertise in team processes and assessment design and four board-certified attending EM physicians with a collective 56 years of experience. The SMEs independently assigned each behavior to a dimension within the taxonomy or left the behavior unassigned for discussion. We calculated percent agreement among SME behavioral classifications for all behaviors. Any behavior with inadequate agreement, defined as < 75% agreement, was discussed. This process was repeated three times until adequate agreement was reached for all dimensions and behaviors. The included behaviors were used to create one or more behaviorally anchored rating scales (BARS) for each leadership dimension. Each scale constitutes one item within the larger measure.

Phase 2: Review and Refinement of the Measure

Content validation followed best practices outlined by Haynes et al.³³ A second group of SMEs, which included eight EM physicians from four different institutions, used a 5-point Likert scale to rate each of the items on: 1) appropriate representation of EM leadership and 2) importance to EM leadership. SMEs were provided with the highest-level behavioral anchor for

each item. A rating of 1 indicated the anchor was not at all representative/important and a rating of 5 indicated that the anchor was highly representative/important. The survey also provided SMEs with the option to provide free-text comments related to the items and anchors. Behavioral anchors with a mean Likert score < 4 in either category, and any associated SME comments, were discussed.

We collected response process validity evidence by evaluating written and verbal feedback from four EM physicians after they piloted the assessment measure on one simulated, and one actual, EM resuscitation. Items identified as confusing, difficult to score, or not representative of EM team leadership were reviewed and modified.

Phase 3: Empirical Testing of the Measure Video Library.

We used a preexisting video library of EM resuscitations for empirical testing and the collection of additional validity evidence. Recordings include both simulated and actual patient care resuscitations. The simulated medical resuscitation scenario and supporting evidence of validity have been previously described.³⁴ Briefly, it involves a patient presenting from a skilled nursing facility with altered mental status and hypoxia that progresses to respiratory distress and cardiac arrest. All teams were interprofessional, with an EM resident physician, a nurse, and a medical student, from one of two different institutions.

The scenario was standardized and included an embedded participant functioning as a nurse. This nurse role was filled by one of two different people who had prior experience as standardized participants and training specific to this scenario. The purpose of this standardized participant was to provide information at set time points (e.g., radiology images), assist with overcoming the artificial aspects of the simulation (e.g., drawing blood), and ensure participant safety (e.g., avoiding contact with the mannequin during defibrillation). Actual patient care videos included the highest-level trauma activations for field responses treated in the emergency department (ED) of a Level I trauma center. Videos were blurred to obscure patient identity.

External Variables for Comparison. Previously collected data associated with the simulated scenarios provided external variables for comparison that we expected to positively correlate with better leadership. These variables included: 1) an 81-item team process (teamwork) score, 2) a 63-item clinical performance (clinical care) score, and 3) a seven-item team level situational awareness score.^{34,35} We also used team leader level of training as an external variable, with the expectation that senior (third- and fourth-year) residents would be better leaders than junior (second year) residents.

Rater Training. Raters included four board-certified EM physicians (EDR, KAJ, MJB, LW), from three academic institutions with experience in graduate medical education (e.g., residency program leadership, simulation program directorship). Raters discussed the assessment measure prior to viewing four training videos representing a range of team leader performance in both simulated and actual resuscitations. The raters met twice to discuss and calibrate scores. We elected not to train to a specific marker of inter-rater reliability (IRR) anticipating that most educators would use this measure with minimal rater training.

Leader Performance Assessment. The team leader assessment measure was used to assess team leader performance in 20 simulated medical resuscitations and 10 actual trauma resuscitations. A sample size of 30 was selected based on available videos and sample sizes from other measure development and validation studies.^{14,17,36} All four trained raters

independently coded each video. Raters were blinded to the previously collected external variable data.

Data Analysis

Phases 1 and 2. Agreement for initial SME categorization of leadership behaviors was determined using percent agreement to identify items that needed further discussion. For the external SME review process, the mean Likert scores and standard deviations (SDs) were calculated for the appropriate representation and the importance of each anchor.

Phase 3. Overall leadership scores were determined by calculating the mean of all items and thus could range from 1 to 5. If an item was scored as not applicable it was removed from the denominator. The IRR for the individual items and the overall score was determined using intraclass coefficients [ICC(C,k)],³⁷ with an ICC of >0.70 considered acceptable.³⁸ Descriptive statistics were used to determine mean ICCs.

For the simulation-based scenarios, mean leadership scores were used to determine the relationship to the external variables expected to correlate with team leadership. Correlation analyses were calculated using Pearson's (*r*) product-moment correlation coefficient. Analyses were performed using R Version 3.5.0 (The R Project).³⁹

RESULTS

Phase 1: Initial Measure Development

We reviewed 11 dimensions in the initial teamwork behavioral taxonomy.³¹ Based on expert consensus a new Communication dimension was added, two dimensions were removed, and three dimensions were combined. The complete list of modifications, the rationale for each modification, and the final six dimensions are provided in the Data Supplement S1 (Tables S1, available as supporting information in the online version of this paper, which is available at <http://onlinelibrary.wiley.com/doi/10.1002/aet2.10321/full>). After removing redundant behaviors, we started with a list of 48 team leadership behaviors. Twenty-seven behaviors were retained for measure development based on SME consensus that these behaviors were: 1) important and relevant to EM resuscitation team leadership and 2) observable. The categorization of the included behaviors and the list of unused behaviors are provided in Data Supplement S1, Tables S1 and S2.

Phase 2: Revised Team Leadership Assessment Measure

Mean SME Likert scores were at least 4 on the 5-point scale for both appropriate representation and importance for all items except Manages Workload (Table 1). The SMEs felt that this behavior was important for EM providers when providing clinical care to an entire ED, but might not be required during a single patient resuscitation. The SMEs also felt that this item would be more important in settings with fewer resources as opposed to large academic centers. Based on these SME comments, and our goal of designing an EM team leadership assessment measure that was widely applicable, we elected to retain this item for testing in Phase 3.

Based on feedback from EM raters we removed behaviors related to summarizing from the monitors progress item and create a new Summarizes Progress item. Additional edits were made to improve readability. The result was a 15-item team leader assessment measure (Data Supplement S1, Appendix S1). Each item is a BARS, and items are grouped within the dimensions identified in Phase 1.

Phase 3: Empirical Data Collection

All four raters scored all 30 observations, representing 29 different team leaders. One participant was the team leader in two different patient care observations. In 13 of the observations (43%) the team leader was

female. In 18 of the observations (60%) the team leader was a junior resident (second year resident).

The overall leadership scores ranged from 2.23 to 4.30 with a mean ($\pm$ SD) score of 3.18 ($\pm$ 0.50). The ICC for the overall team leadership score was 0.79 for all observations, 0.87 for simulation-based observations, and 0.24 for actual patient care observations (Table 2). Within the simulation-based observations, item-level ICCs ranged from 0.29 (Sets Goals) to 0.80 (Maintains Global Perspective), with a mean ($\pm$ SD) of 0.64 ($\pm$ 0.15). Conflict management was scored as not applicable in all observations.

Of the external variables used for comparison, the team leaders' level of training and the team process (teamwork) score showed a weak to moderate correlation with team leader performance; however, they did not rise to the level of statistical significance (Table 3). Team leader performance did not correlate with the team clinical performance or team situational awareness scores.

DISCUSSION

The IRR of our team leadership assessment measure using overall scores, across all observations, met criteria for acceptable reliability, but with a discrepancy between simulated and actual patient care observations. Within the simulation-based scenarios, there were a few items that demonstrated lower IRR (ICC < 0.7) that might benefit from revision. Rater feedback offered insight into modifications that could improve the reliability and relevance of these items. For example, adding a time component to the item Gathers Information may help differentiate excellent behavior by identifying team leaders who prioritize the most important information sources.

Reliability was much lower when applying the BARS to actual patient care observations. Only one item (Delegation: Appropriateness) demonstrated adequate IRR. Variations in leadership were more difficult to discern in these observations due to the increased complexity and variability in actual patient care, including differences in team characteristics (e.g., size), patient condition (e.g., acuity level), and environment (e.g., other patients). Raters also cited preexisting protocols related to trauma care as a potentially limiting factor, noting that in a highly resourced trauma setting, many aspects of care are routinized making it more difficult to detect subtle differences in team leader performance. Increasing the number of

Table 1
Expert Review of the Importance and Appropriate Representation of the Leadership Behavioral Anchors

Item*	Likert Scores, Mean ($\pm$ SD)	
	Appropriate Representation	Importance
Gathers information	4.13 ($\pm$ 0.64)	4.13 ($\pm$ 0.64)
Sets goals	4.50 ($\pm$ 0.76)	4.50 ($\pm$ 0.76)
Facilitates decision making	4.75 ($\pm$ 0.71)	4.75 ($\pm$ 0.46)
Contingency planning	4.38 ($\pm$ 0.74)	4.50 ($\pm$ 0.76)
Role delegation (frequency)	4.63 ($\pm$ 0.52)	4.63 ($\pm$ 0.74)
Role delegation (quality)	4.00 ($\pm$ 0.93)	4.13 ($\pm$ 0.83)
Maintains global perspective	4.38 ($\pm$ 0.74)	4.38 ($\pm$ 0.92)
Resources utilization	4.63 ($\pm$ 0.52)	4.63 ($\pm$ 0.52)
Monitors progress	4.75 ($\pm$ 0.46)	4.75 ($\pm$ 0.46)
Manages team workload	3.88 ($\pm$ 0.83)	4.00 ($\pm$ 0.93)
Explicit communication	4.25 ($\pm$ 0.89)	4.38 ($\pm$ 0.92)
Asserts control	4.38 ($\pm$ 0.46)	4.88 ($\pm$ 0.35)
Promotes team collaboration	4.13 ($\pm$ 0.64)	4.13 ($\pm$ 1.13)
Conflict management	4.25 ($\pm$ 0.89)	4.25 (1.16)

*Only the item is listed here; however, the item and all of the highest-level behavioral anchors were provided for expert review.

Table 2
Total and Item-specific IRR and Rater Feedback for Proposed Leadership Measure

Item	IRR (ICCs)			Feedback From Raters
	Sim	Live	All	
Overall leadership score	0.87	0.24	0.79	
Gathers information	0.39	0.37	0.38	<ul style="list-style-type: none"> Consider adding a timing component (e.g., if easily obtainable information is present but not found or explored for 20 minutes). It's not only if the information is obtained, but also when, that contributes to team efficiency.
Sets goals	0.61	-2.16	0.29	<ul style="list-style-type: none"> Behavior is important and can be scored on both live and video cases. The simulation-based scenarios were easier to score, possibly because they were medical cases. The trauma team resuscitations have protocols that will happen no matter what. General goals with medical resuscitations can go in many directions, which makes this item more relevant and easier to score. More difficult to score in live videos. In trauma settings the goals are more understood by team. Stabilize, diagnose, and often disposition quickly. In medical resuscitations the team is less likely to know the priorities.
Facilitates decisions	0.71	0.57	0.74	<ul style="list-style-type: none"> Behaviors are important and generally easy to recognize.
Contingency planning	0.64	0.35	0.58	<ul style="list-style-type: none"> This was more difficult to score on trauma scenarios, again due to protocols and external decision-making (e.g., surgery team deciding to go to the operating room).
Role delegation: frequency	0.77	-0.02	0.60	<ul style="list-style-type: none"> Behaviors easily identified usually. Sometimes roles are predetermined.
Role delegation: appropriateness	0.72	0.74	0.72	<ul style="list-style-type: none"> Consider combining the two delegation items (frequency and appropriateness) as they assess a similar skill. The anchors for appropriate delegation are not observable unless explicitly asked by the leader. There will be situations in which the leader knows who they are working with and would not need to ask out loud to determine if a team member was capable of a certain task.
Global perspective	0.85	0.60	0.79	<ul style="list-style-type: none"> Behaviors are important and generally easy to recognize.
Resource utilization	0.81	0.50	0.73	<ul style="list-style-type: none"> More easily scored in live videos when there were more people and consultants to organize. In smaller teams, with additional resources unavailable, this was less pertinent. Difficult to understand department's needs/resources at any given moment. May need better definition or anchors or consider combining with managing workload.
Monitors progress	0.63	0.54	0.59	<ul style="list-style-type: none"> Behaviors are important and generally easy to recognize. Most consistently reflected in monitoring the patient's status. Monitoring the team and tasks is also important, but less consistently required. Often times it is obvious when a task is completed (e.g., seeing the blood drawn) or the information was provided without a request from the leader (e.g., a team member announcing the results from ultrasound imaging).
Summarizes progress	0.86	0.64	0.80	<ul style="list-style-type: none"> Behaviors are important and generally easy to recognize. Important to remember that more is not always better. Less effective team leaders sometimes did this more frequently: When they didn't know what to do next they summarized the same information over and over.
Manages workload	0.74	-1.24	0.66	<ul style="list-style-type: none"> Can be challenging to tell if someone is overwhelmed mentally or physically, especially in large teams. More easily scored in smaller teams.
Explicit communication	0.82	0.52	0.74	<ul style="list-style-type: none"> Behaviors are important and generally easy to recognize.
Asserts control	0.77	-0.10	0.64	<ul style="list-style-type: none"> Behaviors are important and generally easy to recognize, although the nature of the team can influence this. There will be instances in which generally effective team leaders struggle to maintain control due to lack of team followership. While team leaders can help facilitate team followership and participation, some of this is dependent on the personality of the team members.
Psychological safety	0.82	-0.71	0.75	<ul style="list-style-type: none"> Behaviors are important and generally easy to recognize. People generally behave professionally, so there will probably be less variability with this score; however, it is important to capture when a team leader does not exhibit these behaviors, even if rare.

(Continued)

Table 2 (continued)

Item	IRR (ICCs)			Feedback From Raters
	Sim	Live	All	
Conflict management	NA	NA	NA	<ul style="list-style-type: none"> Not really utilized as this is uncommon and when it does occur it is often addressed later, rather than at the bedside. Could consider incorporating this into another item as a higher-level behavior, such as psychological safety or asserts control (e.g., if conflict arises the team leader was able to manage it appropriately).

ICC = intraclass correlations; IRR = inter-rater reliability; Sim = simulation.

Table 3

Pearson Correlations Between Team Leadership Scores and Other Variables for the Simulation-based Scenarios

Measure	Level of Training	Team Process	Clinical Performance	Team Situational Awareness
Team leadership ($n = 20$)	$r = 0.25$ $p = 0.30$	$r = 0.39$ $p = 0.09$	$r = 0.15$ $p = 0.64$	$r = 0.00$ $p = 0.99$

r = correlation coefficient; p = probability value.

observations and examining a more diverse set of resuscitations, including medical resuscitations, would provide further insight into the IRR of this measure when applied to actual patient care events.

The reliability of this measure may also be improved with more extensive rater training. Raters represented three institutions; with at least two of the raters having no knowledge of the leaders, the other team members, or the clinical environment. Rater training was intentionally limited, as the measure was intended for EM educators to use without extensive rater training. Results suggested this is feasible for the simulation-based cohort. We suspect that using raters from the same institution would result in even higher IRR, especially for the actual patient care observations, due to a shared understanding of the clinical environment, culture, and team leader expectations. More extensive rater training could also help to build this shared understanding, and this may be necessary to achieve adequate reliability in the clinical setting.

The overall team leadership scores did not significantly correlate with external variables thought to relate to leadership performance. Team leadership, teamwork, and clinical care are all complex constructs, with many confounding patient and team variables. We suspect that the relationship with leadership performance was overshadowed by other factors that contributed to team performance. Another possibility is that the team leadership measure or the external measures do not reflect the desired constructs (e.g., the team leadership measure does not accurately reflect variations in leadership performance). We think that this is less likely given the rigorous development process for these measures. Finally,

the expected relationship between these constructs may not actually exist. There is adequate work supporting the link between team leadership and teamwork, but less is known about the relationship between team leadership and other constructs, such as team situational awareness. These relationships are complex and require consideration of different team and task characteristics. A larger sample size would allow us to examine the relationship between team leadership and other variables while controlling for additional team factors, including team familiarity and aggregate team experience. It is also worth considering other variables that might be expected to improve with higher-quality team leadership that we did not measure, such as team satisfaction or team cohesion.

Future research directions could include revising and reevaluating this measure on a larger sample of actual patient care observations. Specific revisions, such as the examples noted above, would require SME review prior to more widespread testing to evaluate the generalizability of the measure. Additionally, a more rigorous evaluation of the raters' perspective on the measure might provide insight into the challenges of using this measure to assess the complex construct of team leadership. Ideally, EM educators could use this measure to assess EM resident team leaders while on shift. Video-based review allows raters to pause and replay to look for more subtle behaviors and also allows raters to focus on the team leader, without worrying about clinical care or patient safety; however, it may not be practical for resident evaluations. Comparing scores from both live and recorded observations of the same resuscitations would help determine the

appropriateness of using this measure during clinical shifts. While we intended this measure for EM resident education, the content of the measure is not specific to trainees, and we anticipate that a revised version demonstrating acceptable reliability could be used to assess any practitioner functioning as the EM team leader (e.g., attending physicians).

LIMITATIONS

There are several limitations to this study. While we used more encounters than many other observation-based measure development studies,^{14,17,36} the sample size remains a limiting factor when attempting to understand and control for confounding factors. Including actual patient care resuscitations introduced additional team, patient, and environmental variables. The standardized simulation-based scenarios limited some of these variables, but also provided less ecologic validity. We elected to include both actual and simulated patient care scenarios, as both settings are important to assessment in graduate medical education. Including both types of observations also allowed us to address the limitations of the different settings, but required conceding smaller sample sizes for each cohort. Furthermore, the simulation-based observations were all medical resuscitations and the patient care observations were all trauma resuscitations and so the difference in reliability could be attributable the clinical component (trauma vs. medical) or the setting (simulation vs. actual patient care) or both. The SMEs for the measure development process included EM physicians from four different institutions to improve generalizability; however, the team leaders evaluated during actual patient care and simulation came from one and two institutions, respectively. Finally, our actual patient care observations were limited to trauma care and may not have been reflective of all EM resuscitations as a whole.

CONCLUSION

We developed and tested a novel leadership assessment measure for EM resuscitation teams. The validity of the measure is supported by evidence of content validity and response process, as well as acceptable inter-rater reliability when applied to simulated medical resuscitations, but not to actual trauma resuscitations. More observations across a variety of clinical conditions and settings are needed to facilitate additional validity testing.

REFERENCES

1. Kozlowski SW, Gully S, McHugh P, Salas E, Cannon-Bowers J. A dynamic theory of leadership and team effectiveness: Developmental and task contingent leader roles. In: Ferris GR, editor. *Research in Personnel and Human Resources Management*. Greenwich, CT: JAI Press, 1996. pp. 253–306.
2. Neily J, Mills PD, Young-Xu Y, et al. Association between implementation of a medical team training program and surgical mortality. *JAMA* 2010;304:1693–700.
3. Sundstrom E, De Meuse KP, Futrell D. Work teams: applications and effectiveness. *Am Psychol* 1990;45:120.
4. Klein KJ, Ziegert JC, Knight AP, Xiao Y. Dynamic delegation: shared, hierarchical, and deindividualized leadership in extreme action teams. *Admin Sci Quart* 2006;51:590–621.
5. Sebok-Syer SS, Chahine S, Watling CJ, Goldszmidt M, Cristancho S, Lingard L. Considering the interdependence of clinical performance: implications for assessment and entrustment. *Med Educ* 2018;52:970–80.
6. Blumenthal DM, Bernard K, Bohnen J, Bohmer R. Addressing the leadership gap in medicine: residents' need for systematic leadership development training. *Acad Med* 2012;87:513–22.
7. Jardine D, Correa R, Schultz H, et al. The need for a leadership curriculum for residents. *J Grad Med Educ* 2015; 7:307–9.
8. Goldman E, Plack M, Roche C, Smith J, Turley C. Learning clinical versus leadership competencies in the emergency department: strategies, challenges, and supports of emergency medicine residents. *J Grad Med Educ* 2011;3:320–25.
9. Rosenman ED, Branzetti JB, Fernandez R. Assessing team leadership in emergency medicine: the milestones and beyond. *J Grad Med Educ* 2016;8:332–40.
10. 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; 90:1408–22.
11. Cooper S, Wakelam A. Leadership of resuscitation teams: "lighthouse leadership". *Resuscitation* 1999;42:27–45.
12. Youngblood P, Harter PM, Srivastava S, Moffett S, Heinrichs WL, Dev P. Design, development, and evaluation of an online virtual emergency department for training trauma teams. *Simul Healthc* 2008;3:146–53.
13. Sakran JV, Finneman B, Maxwell C, et al. Trauma leadership: does perception drive reality? *J Surg Educ* 2012; 69:236–40.
14. Nadkarni LD, Roskind CG, Auerbach MA, Calhoun AW, Adler MD, Kessler DO. The development and validation of a concise instrument for formative assessment of team leader performance during simulated pediatric resuscitations. *Simul Healthc* 2018;13:77–82.
15. Frankel A, Gardner R, Maynard L, Kelly A. Using the communication and teamwork skills (CATS) assessment

- to measure health care team performance. *Jt Comm J Qual Patient Saf* 2007;33:549–58.
16. Accreditation Council for Graduate Medical Education, American Board of Emergency Medicine. The Emergency Medicine Milestone Project. July 2015. Available at: <http://www.acgme.org/acgmeweb/Portals/0/PDFs/Milestones/EmergencyMedicineMilestones.pdf>. Accessed on July 10, 2017
 17. Walker S, Brett S, McKay A, Lambden S, Vincent C, Sevdalis N. Observational skill-based clinical assessment tool for resuscitation (OSCAR): development and validation. *Resuscitation* 2011;82:835–44.
 18. Abell N, Springer DW, Kamata A. Developing and Validating Rapid Assessment Instruments. New York, NY: Oxford University Press, 2009.
 19. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;119(166):e7–16.
 20. Messick S. Validity. In: Linn RL, editor. *Educational Measurement*, 3rd ed. New York, NY: Macmillan, 1988.
 21. Rosenman ED, Shandro JR, Ilgen JS, Harper AL, Fernandez R. Leadership training in health care action teams: a systematic review. *Acad Med* 2014;89:1295–306.
 22. Baker VO, Cuzzola R, Knox C, et al. Teamwork education improves trauma team performance in undergraduate health professional students. *J Educ Eval Health Prof* 2015;12:36.
 23. Cooper S, Cant R, Connell C, et al. Measuring teamwork performance: validity testing of the Team Emergency Assessment Measure (TEAM) with clinical resuscitation teams. *Resuscitation* 2016;101:97–101.
 24. Watkins SC, Roberts DA, Boulet JR, McEvoy MD, Weiniger MB. Evaluation of a simpler tool to assess nontechnical skills during simulated critical events. *Simul Healthc* 2017;12:69–75.
 25. Ford K, Menchine M, Burner E, et al. Leadership and teamwork in trauma and resuscitation. *West J Emerg Med* 2016;17:549.
 26. Hargett CW, Doty JP, Hauck JN, et al. Developing a model for effective leadership in healthcare: a concept mapping approach. *J Healthc Leadership* 2017;9:69.
 27. Leenstra NF, Jung OC, Johnson A, Wendt KW, Tulleken JE. Taxonomy of trauma leadership skills: a framework for leadership training and assessment. *Acad Med* 2016;91:272–81.
 28. Robinson PS, Shall E, Rakhit R. Cardiac arrest leadership: in need of resuscitation? *Postgrad Med J* 2016;92:715–20.
 29. Taylor KL, Parshuram CS, Ferri S, Mema B. A description of the “event manager” role in resuscitations: a qualitative study of interviews and focus groups of resuscitation participants. *J Crit Care* 2017;39:254–8.
 30. Chiu CJ. Development and Validation of Performance Assessment Tools for Interprofessional Communication and Teamwork (PACT) [dissertation]. Seattle: University of Washington, 2014.
 31. Marks MA, Mathieu JE, Zaccaro SJ. A temporally based framework and taxonomy of team processes. *Acad Manage Rev* 2001;26:356–76.
 32. Hsu CC, Sandford BA. Delphi technique: making sense of consensus. *Pract Assess Res Eval* 2010;12(10).
 33. Haynes SN, Richard D, Kubany ES. Content validity in psychological assessment: A functional approach to concepts and methods. *Psychol Assessment* 1995;7:238.
 34. Rosenman ED, Dixon AJ, Webb JM, et al. A simulation-based approach to measuring team situational awareness in emergency medicine: a multicenter, observational study. *Acad Emerg Med* 2018;25:196–204.
 35. Fernandez R, Pearce M, Grand JA, et al. Evaluation of a computer-based educational intervention to improve medical teamwork and performance during simulated patient resuscitations. *Crit Care Med* 2013;41:2551–62.
 36. Andersen PO, Jensen MK, Lippert A, Østergaard D, Klausen TW. Development of a formative assessment tool for measurement of performance in multi-professional resuscitation teams. *Resuscitation* 2010;81:703–11.
 37. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1996;1:30.
 38. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assessment* 1994;6:284.
 39. The R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>. Accessed August 8, 2018.

Supporting Information

The following supporting information is available in the online version of this paper available at <http://onlinelibrary.wiley.com/doi/10.1002/aet2.10321/full>

Data Supplement S1. Supplemental material.