An Event-based Approach to Measurement: Facilitating Observational Measurement in Highly Variable Clinical Settings

Rosemarie Fernandez, MD¹ , Elizabeth D. Rosenman, MD², Sarah Broliar², Anne K. Chipman, MD, MS², Colleen Kalynych, MSH, EdD³, Marie C. Vrablik, MD, MCR², Joseph R. Keebler, PhD⁴, and Elizabeth H. Lazzara, PhD⁴

ABSTRACT

Background: Translational research in medical education requires the ability to rigorously measure learner performance in actual clinical settings; however, current measurement systems cannot accommodate the variability inherent in many patient care environments. This is especially problematic in emergency medicine, where patients represent a wide spectrum of severity for a single clinical presentation. Our objective is to describe and implement EBAM, an event-based approach to measurement that can be applied to actual emergency medicine clinical events.

Methods: We used a four-step event-based approach to create an emergency department trauma resuscitation patient care measure. We applied the measure to a database of 360 actual trauma resuscitations recorded in a Level I trauma center using trained raters. A subset ($n = 50$) of videos was independently rated in duplicate to determine inter-rater reliability. Descriptive analyses were performed to describe characteristics of resuscitation events and Cohen's kappa was used to calculate reliability.

Results: The methodology created a metric containing both universal items that are applied to all trauma resuscitation events and conditional items that only apply in certain situations. For clinical trauma events, injury severity scores ranged from 1 to 75 with a mean ($\pm$ SD) of 21 ($\pm$ 15) and included both blunt (254/360; 74%) and penetrating (86/360; 25%) traumatic injuries, demonstrating the diverse nature of the clinical encounters. The mean ($\pm$ SD) Cohen's kappa for patient care items was 0.7 ($\pm$ 0.3).

From the ¹Department of Emergency Medicine and the Center for Experiential Learning and Simulation, College of Medicine, University of Florida, Gainesville, FL; the ²Department of Emergency Medicine, University of Washington, Seattle, WA; the ³Department of Emergency Medicine, Office of Educational Affairs, University of Florida College of Medicine—Jacksonville, Jacksonville, FL; and the ⁴Department of Human Factors and Behavioral Neurobiology, College of Arts and Sciences, Embry-Riddle Aeronautical University, Daytona Beach, FL.

Received June 2, 2019; revision received September 9, 2019; accepted September 11, 2019.

Funding and support for this project was provided by the Agency for Healthcare Research and Quality (1R18HS022458-01A1 [RF]) and the Department of Defense Congressionally Directed Medical Research Program (W81XWH1810089 [RF]). The funding sources 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.

Conflict of interest: RF's institution has received grant money from the Agency for Healthcare Research and Quality, the Department of Defense, and the Washington State Department of Labor and Industries to conduct research conceived and written by RF. RF reports personal payment from Physio-Control, Inc. for speaker fees. EDR's institution has received grant money from the Agency for Healthcare Research and Quality and the Department of Defense to conduct research conceived and written by RF. EDR reports personal payment from Physio-Control, Inc. for speaker fees. SB's institution has received grant money from the Agency for Healthcare Research and Quality and the Department of Defense to conduct research conceived and written by RF. AKC reports no conflict of interest. CK's institution has received grant money from the Agency for Healthcare Research and Quality and the Department of Defense to conduct research conceived and written by RF. JRK reports no conflict of interest. EHL reports no conflict of interest. MCV reports no conflict of interest.

Author contributions: concept and design—RF, EDR, JRK, and EHL; acquisition, analysis, or interpretation of data—RF, EDR, SB, CK, AKC, and MCV; drafting of manuscript—RF and EDR; critical revision of the manuscript for important intellectual content—RF, EDR, CK, AKC, MCV, SB, JRK, and EHL; statistical expertise—JRK and EHL; and obtained funding—RF.

Supervising Editor: Sally Santen, MD, PhD.

Address for correspondence and reprints: Rosemarie Fernandez MD; e-mail: fernandez.r@ufl.edu.

AEM EDUCATION AND TRAINING 2020;4:147–153.

Conclusion: We present an event-based approach to performance assessment that may address a major gap in translational education research. Our work centered on assessment of patient care behaviors during trauma resuscitation. More work is needed to evaluate this approach across a diverse array of clinical events.

Medical education research aims to improve physician training, thus making the delivery of health care safer and more effective. As noted by McGaghie,¹ the “downstream goals of medical education research are to demonstrate that educational interventions contribute to physician competence measured in the classroom, educational laboratory, and patient care setting.” This suggests an important role for translational medical education research, where testing of interventions often begins with knowledge assessments and performance in simulated environments (T1) and advances to assessments of actual clinical performance (T2) before demonstrating impact on patient outcomes (T3).

There is a gap in evaluating the impact of educational interventions on clinical performance. Most commonly, assessments target knowledge and skill acquisition, without evaluating whether these skills translate into behavioral change and improved clinical care.² Measuring individual skills, trained behaviors, and team performance in the clinical setting are key components of effective translational research in medical education;³ however, the development of metrics remains a limitation.⁴ It is challenging to create a behavioral checklist that is reliable, can discriminate different levels of performance, and can accommodate the variability encountered in the actual clinical environment. Measures that

perform well in the simulated setting do not necessarily translate to clinical use. Measures intended for clinical care may be too general to discriminate performance quality or may rely too heavily on subject matter expertise for accurate scoring.⁵⁻⁷

Highly variable tasks, such as those commonly present in emergency medicine, pose further challenges to measuring learner performance during actual clinical events.⁸ In the emergency department (ED), patients represent a wide range of complexity and potential clinical instability within a single diagnosis. In such situations, variability in the patient’s condition and the clinical environment significantly impacts which behaviors are indicated and in what order. We present an event-based approach to measurement (EBAM) that allows for the measurement of complex behavioral processes across highly variable clinical events using trauma team resuscitations as an example.

METHODS

EBAM

Event-based approach to measurement adapts the concepts used in event-based training for clinical observation-based measurement design.⁹ Event-based training follows a conceptual design process that supports the

Figure 1. The event-based approach to measurement (EBAM) process.

systematic introduction of work-related tasks to purposefully elicit the behaviors or competencies that are the focus of training interventions.⁹ We translated this concept to develop EBAM, a four-step process (Figure 1) that allows for the development of measures that are 1) modifiable for different patient conditions and tasks, 2) reliable and supported by validity evidence,¹⁰ and 3) directly linked to evidence-based practice. As with event-based training, the core content of EBAM is based on identification of clinical events and triggers. In event-based training the triggers are predefined and controlled, which ensures that the learner has an opportunity to reach all learning objectives. In EBAM, the triggers are dependent on clinical factors (e.g., patient condition, team behaviors), which ensures the assessment only includes pertinent items. This can be applied to individual or team-level performance depending on the research target.

We applied EBAM for the purpose of evaluating team-level clinical care during trauma resuscitations.

While Advanced Trauma Life Support (ATLS) provides general guidelines for trauma care, the execution of trauma resuscitation is much more complex and dependent on highly variable patient factors. There are universal events that contain behaviors that should be performed for every trauma regardless of the patient, the etiology of the trauma, or environmental factors. However, there are other conditional behaviors that depend on the patient's state, for example, the presence of shock physiology. A "shock" event would contain behavioral responses expected for a patient in clinical shock, but would not be expected when caring for a hemodynamically stable patient. Together, the universal and shock behavioral responses then become measurement items that can flex to accommodate both hemodynamically stable and unstable patients. In Figure 2, we provide an example that illustrates how EBAM can handle the variable need for procedural tasks during trauma resuscitations.

Figure 2. Example of event-based measurement system (event + trigger + measurement item) In this example, there are several triggers (T1 – T5), either alone or in combination, that should prompt the behavioral responses and thus measurement items, listed in the event *Procedure: Thoracostomy*. Initiating the *Procedure: Thoracostomy* also activates a second, related event: *Procedure: Universal*. This event contains items universally used during major resuscitation-related procedures. It would be expected that other procedures, such as central venous catheter placement would also trigger the *Procedure: Universal* event. Abbreviations: CXR = chest x-ray; CVC = central venous catheter; PTX = pneumothorax; E-FAST = extended focused assessment with sonography in trauma; SBP = systolic blood pressure

Study Design and Setting

Trauma resuscitation events ($n = 360$) were recorded at Harborview Medical Center, an urban, Level I trauma center in the University of Washington health care system. The University of Washington Human Subjects Division approved this study.

Trauma Resuscitation Events

Eligible resuscitations included all adult patients presenting to the ED meeting American College of Surgeons recommendations and Harborview Medical Center institutional criteria for trauma evaluation.^{11,12} Prisoners and pregnant women were excluded. Patient characteristics were obtained from the Harborview Medical Center trauma registry.

Trauma Team Patient Care Measure Development

Step 1: Incorporate Evidence-based Practice. Step 1 focused on facilitating the inclusion of evidence-based practices for trauma management during measure development. We identified published adult trauma patient care checklists that have been applied to both simulated¹³ and live^{14–17} patient care events, as well as standards of trauma care, including ATLS.¹⁸ Subject matter experts (SMEs), four board-certified emergency medicine physicians and one clinical nurse, reviewed the guidelines and assessment measures to identify items that 1) were appropriate for trauma resuscitations, 2) apply across all types and severity of trauma resuscitations (universal items), and 3) were indicated in certain clinical presentations but were not universally relevant (conditional items). For conditional items the SMEs determined under what conditions they should be performed. Finally, SMEs noted which items were time-sensitive. All behavioral and time-based items, universal and conditional, were included in a single patient care behavioral measure.

Step 2: Define Events. In Step 2, we defined and described trauma resuscitation behavioral events. Universal items from Step 1 were grouped into events expected to occur during every trauma resuscitation, such as “performs a primary survey.” The conditional items were also placed into events as appropriate. For example, the event “shock, presumed hemorrhagic” included items like “initiates blood transfusion.” SMEs reviewed all universal and conditional events and items to ensure that the items were appropriately assigned.

Step 3: Define Event Triggers. Step 3 focused on defining the clinical triggers that prompt the conditional events. If an event is “triggered,” then behaviors in that event should be included in the final performance assessment. The investigators worked with SMEs to identify observable clinical triggers. We prioritized triggers that were as specific as possible. For example, the item “initiates blood transfusion,” as mentioned above, is within the event “shock, presumed hemorrhagic” and is triggered by two or more low blood pressures (systolic blood pressure < 90 mm Hg). Other possible triggers, such as elevated heart rate, may occur for multiple reasons (e.g., pain) and thus were not considered specific enough to serve as a trigger for initiating a blood transfusion. We also had to limit triggers to items that could be reliably observed on video. A low hematocrit may prompt a blood transfusion; however, this laboratory value was not easily observable. All event triggers were reviewed by SMEs after they had observed 10 video-recorded resuscitations to ensure that triggers met the criteria of being 1) observable, 2) clinically appropriate, and 3) specific.

Step 4: Test and Refine Measure. In Step 4 the measure was tested to ensure that proposed behaviors, events, and triggers could be reliably observed. Preliminary determination of reliability during measure development informed ongoing rater training and prompted modifications to the EBAM items that were confusing, poorly defined, or assigned to an inappropriate event. Initial ratings were performed as a group with a “think-aloud” approach to ensure that important behavioral items and events were captured to the best of our ability.¹⁹

Steps 1 through 4 resulted in a list of trauma resuscitation patient care items (Data Supplement S1, Table S1, available as supporting information in the online version of this paper, which is available at <http://onlinelibrary.wiley.com/doi/10.1002/aet2.10395/full>) that included universal and conditional behaviors, behavioral triggers, and the time to key behaviors. The involvement of SMEs and clinical guidelines (ATLS) helped to establish evidence of content validity. These items were used to code recorded trauma resuscitations. Trigger variables were then used during data analysis to identify which items were appropriate for each individual trauma resuscitation. The total possible points (denominator) ranged from 20 to 38 based on the number of conditional items deemed appropriate for that particular patient resuscitation.

Data Coding

Coding was performed using Noldus Observer[®] XT (Leesburg, VA) software. Primary coders were part-time research assistants who also worked as hospital volunteers in high-acuity patient care settings (ED and pediatric intensive care unit) but who were not health care providers. A subset of resuscitations was coded in duplicate by a board-certified emergency physician. Additionally, board-certified emergency medicine physicians coded in duplicate all items requiring clinical judgment. The overall coding approach maximized the use of nonclinical experts and limited the coding burden for the physician investigators by using nonclinical coders for all scored items and for the triggers that did not require clinical judgment.²⁰ An example of an item requiring clinical judgment was “vasopressors indicated,” which triggered the conditional item “team orders vasopressor infusion.” An example of a trigger not requiring clinical judgment was “two consecutive low blood pressures,” which triggered the conditional item “team initiates blood transfusion.” Additional examples are provided in Data Supplement S1, Table S1. All coders trained until they reached a Cohen’s $\kappa > 0.75$ across a range of performance episodes that varied with regard to complexity, illness severity, and patient care requirements. Rater training represents evidence of response process validity.²¹ To determine inter-rater reliability, 14% of resuscitations (50/360) were coded in duplicate by one of the investigators (LR, RF). Disagreements were reviewed by the coders and reconciled. Inter-rater reliability represents evidence of internal structure validity.²¹

Data Analysis

Statistical analyses were performed using SPSS Statistics version 19 (IBM Corp.) and the open-source statistical program R version 3.5. We calculated Cohen’s kappa to determine the degree of inter-rater reliability on trauma resuscitation patient care measure items.²² We computed descriptive statistics (mean and standard deviation [SD]) for all resuscitation patient characteristics.

RESULTS

From March 2016 through February 2018, we recorded 360 trauma resuscitation events and were able to match 342 with descriptive characteristics in the Harborview Medical Center trauma registry. All 360 resuscitations were coded for trauma resuscitation patient care items. Overall inter-rater reliability across

all items was Cohen’s $\kappa = 0.7$ (0.3). Characteristics of the trauma resuscitations are described in Data Supplement S1, Table S2, and demonstrate a wide range of clinical events, with injury severity scores ranging from 1 to 75 with a mean ($\pm$ SD) of 21 ($\pm$ 5).

DISCUSSION

In this article we present an approach to observational measurement for research that captures granular metrics while accommodating high levels of clinical variability. Such granularity can be important when assessing behaviors and actions in medical education–based clinical trials. It is important that specific aspects of training can be directly linked with specific behaviors. Current assessment tools meant for general resident assessment (e.g., direct observation of procedural skills [DOPS]²³ and mini-Clinical Evaluation Exercise [mini-CEX]²⁴) capture global performance but cannot fully explicate changes in performance at a highly detailed level.

The example used in this article focuses on assessment of trauma resuscitative care. However, EBAM may be applicable to other topics, including other areas of team-based clinical care (e.g., medical resuscitations), procedures, and interpersonal skills (e.g., managing conflict or delivering bad news). The strength of EBAM is its ability to handle patient variability in a systematic way. Thus, undifferentiated or unstable patients (e.g., rapid response clinical events), rapidly changing clinical environments (e.g., disaster settings or low-resourced settings), and highly complex patient care events are well suited to EBAM. While we evaluated team performance, we believe that this same approach could be used for individual assessments. The ability to use event-based training for simulation-based research and then transform this work into EBAM for clinical observation provides a mechanism for translational research in graduate medical education and interdisciplinary training.

Limitations

While our approach addresses some of the challenges in observational measurement, there are important limitations to recognize. We presented reliability for the entire measure and not at the item level. This was done because there was considerable variability regarding the number of times certain items were indicated and kappa is sensitive to prevalence.²⁵ Reliability for rarely occurring items may not be accurately reflected. We were able to use video recordings of trauma resuscitations, thus enabling coders to watch the same behaviors

multiple times to help ensure accurate coding. EBAM could be applied to live observations as well; however, one might have to sacrifice the detail necessary for high-level research, especially in chaotic environments such as resuscitations. Further research would be needed to determine how the methodology could be best structured to assess live events. Additionally, both event-based training and EBAM focus on observable behaviors. Thus, to accurately measure cognitive skills, there must be acceptable behavioral proxies that accurately represent the cognitive construct of interest. Finally, EBAM is resource-intensive. To adapt to patient variability, the measure must consider a number of conditional events and triggers and how these connect. As a result the test/refine process can be more time-consuming than what is needed with simulation-based measure development. While we acknowledge this resource burden, we feel EBAM provides a mechanism for the highly detailed measurement necessary for research, thus justifying the extra effort.

CONCLUSIONS

In conclusion, we describe an approach to behavioral assessment that addresses a major barrier to medical education-based translational research. Additional psychometric testing is needed to evaluate the application of this technique to other clinical events.

References

- McGaghie WC. Medical education research as translational science. *Sci Transl Med* 2010;2:19cm8.
- Prystowsky JB, Bordage G. An outcomes research perspective on medical education: the predominance of trainee assessment and satisfaction. *Med Educ* 2001;35:331–6.
- McGaghie WC, Draycott TJ, Dunn WF, Lopez CM, Stefanidis D. Evaluating the impact of simulation on translational patient outcomes. *Simul Healthc* 2011;6(Suppl): S42–47.
- McGaghie WC, Issenberg SB, Barsuk JH, Wayne DB. A critical review of simulation-based mastery learning with translational outcomes. *Med Educ* 2014;48:375–85.
- Whelan GP, Boulet JR, McKinley DW, et al. Scoring standardized patient examinations: lessons learned from the development and administration of the ECFMG Clinical Skills Assessment (CSA (R)). *Med Teach* 2005;27:200–6.
- Holmboe ES. Faculty and the observation of trainees' clinical skills: problems and opportunities. *Acad Med* 2004;79:16–22.
- Grand JA, Pearce M, Rench TA, Chao GT, Fernandez R, Kozlowski SW. Going DEEP: guidelines for building simulation-based team assessments. *BMJ Qual Saf* 2013;22:436–48.
- Norcini JJ. Current perspectives in assessment: the assessment of performance at work. *Med Educ* 2005;39:880–9.
- Fowlkes J, Dwyer DJ, Oser RL, Salas E. Event-based approach to training (EBAT). *Int J Aviat Psychol* 1998;8:209–21.
- Messick S. Validity. In: Linn RL, editor. *Educational Measurement*. New York: American Council on Education and Macmillan, 1989:13–103.
- Committee on Trauma. *Resources for Optimal Care of the Injured Patient*. Chicago, IL: American College of Surgeons, 2014.
- Washington State Department of Health Office of Community Health Systems - Emergency Medical Services & Trauma Section. *Trauma Clinical Guideline: Trauma Team Activation Criteria*. 2016. Available at: <http://provideresource.uwmedicine.org/flexpaper/trauma-team-activation-criteria>. Accessed October 2, 2018.
- Holcomb JB, Dumire RD, Crommett JW, et al. Evaluation of trauma team performance using an advanced human patient simulator for resuscitation training. *J Trauma* 2002;52:1078–85.
- Lubbert PH, Kaasschieter EG, Hoorntje LE, Leenen LP. Video registration of trauma team performance in the emergency department: the results of a 2-year analysis in a Level 1 trauma center. *J Trauma* 2009;67:1412–20.
- Ritchie PD, Cameron PA. An evaluation of trauma team leader performance by video recording. *Aust N Z J Surg* 1999;69:183–6.
- Sugrue M, Seger M, Kerridge R, Sloane D, Deane S. A prospective study of the performance of the trauma team leader. *J Trauma* 1995;38:79–82.
- Kelleher DC, Bose RJ, Waterhouse LJ, Carter EA, Burd RS. Effect of a checklist on advanced trauma life support workflow deviations during trauma resuscitations without pre-arrival notification. *J Am Coll Surg* 2014;218:459–66.
- Committee on Trauma. *Advanced Trauma Life Support for Doctors: Student Course Manual* 10th ed. Chicago, IL: American College of Surgeons, 2018.
- Pinnock R, Fisher TL, Astley J. Think aloud to learn and assess clinical reasoning. *Med Educ* 2016;50:585–6.
- 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.
- Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;119:166.e7–16.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.

23. Norcini JJ, McKinley DW. Assessment methods in medical education. *Teach Teach Educ* 2007;23:239–50.
24. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. *Ann Intern Med* 2003;138:476–81.
25. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992;304:1491.

Supporting Information

The following supporting information is available in the online version of this paper available at <http://onlinelibrary.wiley.com/doi/10.1002/aet2.10395/full>

Data Supplement S1. Supplemental material.