Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Establishing Objective Measures of Clinical Competence in Undergraduate Medical Education Through Immersive Virtual Reality

Matthew W. Zackoff, MD, MEd; Daniel Young, MD; Rashmi D. Sahay, MD, MS; Lin Fei, PhD; Francis J. Real, MD, MEd; Amy Guiot, MD, MEd; Corinne Lehmann, MD, MEd; Melissa Klein, MD, MEd

From the Department of Pediatrics, University of Cincinnati College of Medicine (MW Zackoff, L Fei, FJ Real, A Guiot, C Lehmann, M Klein), Cincinnati, Ohio; Division of Critical Care Medicine, Department of Pediatrics, Cincinnati Children's Hospital Medical Center (MW Zackoff), Cincinnati, Ohio; Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati Children's Hospital Medical Center (D Young), Cincinnati, Ohio; Division of Biostatistics and Epidemiology, Department of Pediatrics, Cincinnati Children's Hospital Medical Center (RD Sahay, L Fei), Cincinnati, Ohio; Division of General and Community Pediatrics, Department of Pediatrics, Cincinnati Children's Hospital Medical Center (FJ Real, M Klein), Cincinnati, Ohio; Division of Hospital Medicine, Department of Pediatrics, Cincinnati Children's Hospital Medical Center (A Guiot, M Klein), Cincinnati, Ohio; and Division of Adolescent Medicine, Department of Pediatrics, Cincinnati Children's Hospital Medical Center (C Lehmann), Cincinnati, Ohio

The authors have no conflicts of interest to declare.

Address correspondence to Matthew W. Zackoff, MD, MEd, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, MLC 2005, Cincinnati, OH 45229 (e-mail: matthew.zackoff@cchmc.org).

Received for publication June 10, 2020; accepted October 17, 2020.

Abstract

Objective: The Association of American Medical Colleges defines recognition of the need for urgent or emergent escalation of care as a key Entrustable Professional Activity (EPA) for entering residency (EPA#10). This study pilots the use of an immersive virtual reality (VR) platform for defining objective observable behaviors as standards for evaluation of medical student recognition of impending respiratory failure.

Methods: A cross-sectional observational study was conducted from July 2018 to December 2019, evaluating student performance during a VR scenario of an infant in impending respiratory failure using the OculusRift VR platform. Video recordings were rated by 2 pair of physician reviewers blinded to student identity. One pair provided a consensus global assessment of performance (not competent, borderline, or competent) while the other used a checklist of observable behaviors to rate performance. Binary discriminant analysis was used to identify the observable behaviors that predicted the global assessment rating.

Results: Twenty-six fourth year medical students participated. Student performance of 8 observable behaviors was found to be most predictive of a rating of competent, with a 91% probability. Correctly stating that the patient required an escalation of care had the largest contribution toward predicting a rating of competent, followed by commenting on the patient's increased heart rate, low oxygen saturation, increased respiratory rate, and stating that the patient was in respiratory distress.

Conclusions: This study demonstrates that VR can be used to establish objective and observable performance standards for assessment of EPA attainment − a key step in moving towards competency based medical education.

Keywords: clinical assessment; competence; respiratory distress; virtual reality

Academic Pediatrics 2020;XXX:1−5

What's New?

Immersive virtual reality (VR) was successfully used as a platform to establish competency standards for the assessment of pediatric respiratory distress. The VR platform delineated key observable behaviors that correlated with faculty ratings of global level of competence.

Competency-based medical education (CBME) is anchored in the concept that trainees master skills at different paces. Progression through the educational continuum should be dictated by demonstrating proficiencies required for transition to the subsequent rank.[1] The Association of American Medical Colleges (AAMC) published Core Entrustable Professional Activities (EPAs) for entering residency which describes specific skills and behaviors expected of all graduating medical students upon entering their first day of residency.[2] However, before CBME can transition from promise to practice, objective measures to assess performance are required.[3]

Simulation-based medical education (SBME) offers students a safe environment to perform skills and has demonstrated improved educational outcomes compared to traditional didactics.[4−6] While standardized patient encounters, a form of SBME, have become the gold standard for clinical skills assessment, their application to the array of clinical competencies remains limited.[7,8]

Specifically, Core EPA #10 for students entering residency requires students to demonstrate recognition of patients requiring urgent or emergent care.[2] In pediatrics, respiratory distress from bronchiolitis is the most common cause of hospitalization for infants, with nearly 14% of hospitalized patients progressing to respiratory failure.[9] Unfortunately, standardized patients for pediatric respiratory distress are not on option, and many available patient simulators cannot display several critical exam findings (eg, mental status, work of breathing) needed to create realistic conditions for an accurate assessment of competency.

Immersive virtual reality (VR) simulation is a promising new approach to SBME, whereby students are taken to the patient's bedside within a virtual 3D environment. VR simulations promote deliberate practice[10] of skills through safe and realistic interactions with graphical character representatives (avatars). VR has successfully been used for training in various contexts, such as performing procedures,[11] learning empathy,[12] addressing vaccine hesitancy,[13,14] and performing a clinical assessment.[15–17] However, VR has yet to be leveraged for the establishment of competency standards or formal assessment of performance. To address this gap, our study aimed to establish competency standards related to student recognition of impending respiratory failure using an immersive VR platform, using the clinical scenario of an infant admitted with bronchiolitis.

Materials and Methods

Setting and Study Population

A cross-sectional observational study was conducted at Cincinnati Children's Hospital Medical Center in association with the University of Cincinnati College of Medicine, from July 2018 to December 2019. Fourth-year medical students were recruited via email and were provided a $20 gift card for their voluntary participation. Consent was obtained per our Institutional Review Board's approval.

Curriculum Design

The VR scenario's development and content including a simulated inpatient environment with virtual patient and preceptor avatars, vital signs monitor, and room décor along with functionality (visual and auditory cues including patients' breath sounds) has been previously described in *Academic Pediatrics*[15] (https://drive.google.com/file/d/1m-1j7hbxvIu-dK1jdgz9MRQYcubS6-IS/view?usp=sharing) with demonstrated effectiveness as a teaching tool.[16] To establish our competency standards, we focused on a case of impending respiratory failure during which the virtual infant displayed altered mental status, increased work of breathing, abnormal breath sounds, tachycardia, tachypnea, and hypoxia—consistent with a need for escalation of clinical care.[15,16]

Following orientation to the VR environment and functionality, the student was provided a prompt with the pertinent history and presumptive diagnosis of viral bronchiolitis. The student was asked to verbally report the physical exam findings and interpretation of vital signs to the avatar preceptor, provide an overall assessment of the patient's clinical status, and describe next steps for management. If the student did not independently state whether the patient required an escalation of care, the student was asked by the avatar preceptor, "Do you think the patient is stable for the floor?" Following completion of the session, students were provided feedback by a study author (M.Z.) on overall performance. The session lasted approximately 20 minutes and concluded with a demographic survey.

Standard Setting Approach

VR sessions were video recorded, deidentified, and stored on an internal password protected drive to facilitate review. Our approach for establishing standards of competence was based on the borderline group method, a strategy previously utilized for standardized patient encounters.[18,19] This methodology involves cross-referencing 2 assessment strategies to establish consistent criteria for "passing" 1) a categorized global assessment of performance (competent, borderline, or not-competent) and 2) performance on an itemized observable behavior checklist.

Two physicians with masters training in education and expertise in medical student education and evaluation through their roles as pediatric student clerkship directors (A.G., C.L.) performed a blinded independent review of each student's video session and provided the global assessment of performance. There was no predetermined description of the global assessment groups, and the behavior checklist was not provided to minimize bias. However, both reviewers were prompted to consider the AAMC core EPAs for entering residency when performing their assessment of the student. The reviewers met after completing independent review of batches of 5 recordings to discuss any discrepancies and reach an overall consensus score for each student. This was to ensure ongoing calibration in scoring and to provide a consensus global assessment for standard setting. Generally, reviewers agreed on the consensus score and discrepancies were rarely identified. The videos were also reviewed by a second group of physicians (F.R., D.Y.) using a structured observable behavior checklist (Fig. 1), which had been developed for a previous VR study using a modified Delphi approach.[16] Two sample scenarios were graded independently by the 2 reviewers, followed by a debriefing session to compare scores and reach consensus to enhance reliability. A key grading perspective established during this debriefing was that students who required prompting to state that the patient required an escalation of care would still receive credit for a correct response due to having an accurate interpretation of the clinical scenario. Reviewer 1's scores were used for standard setting while Reviewer 2's scores were used to assess interrater reliability. Data was entered into a secure web-based application (Research Electronic Data Capture).[20]

	REPORTER Specific Findings Reported	INTERPRETER Interpretation of Findings	
Mental Status			
Alertness	☐ Eyes closed ☐ Not awake and/or not alert	☐ Altered mental status and/or	
Activity Level	☐ No movement and/or no activity	☐ Lethargic	
Work of Breathing			
Head Bobbing/Flaring	☐ Head bobbing	☐ Increased work of breathing	
Retractions	☐ Suprasternal ☐ Subcostal		
Belly Breathing	☐ Belly breathing		
Breath Sounds			
Aeration	☐ Audible throughout	☐ Obstructive lung disease and/or	
I/E ratio	☐ Prolonged expiratory phase		
Quality	☐ Coarse and/or wheezing	☐ Abnormal	
Vital Signs			
Respiratory Rate	☐ ~60 breaths per minute	☐ Tachypnea and/or fast	
Oxygen Saturation	☐ Saturation oof ~92%	☐ Hypoxemia and/or low	
Heart Rate	☐ ~160 beats per minute	☐ Tachycardia and/or fast	
ADDITIONAL COMPONENTS			
Overall Assessment Correct?			
Respiratory distress **WITH** impending respiratory failure		☐ Yes	☐ No
Proposed Management			
Recognize need for urgent/emergent care (i.e. transfer to ICU, high flow nasal cannula)		☐ Yes	☐ No

Figure 1. *Checklist of observable behaviors for the impending respiratory failure case scenario.*

A key component of the borderline group approach is the use of an itemized observable behavior checklist. However, due to the complexity of performing a clinical assessment, the list of potential observable behaviors that take place are extensive, ranging from reporting and interpreting individual findings through synthesizing information into an overall assessment. We utilized binary discriminant analysis to identify which observable behaviors best predicted the global assessment of performance, allowing the creation of a core set of observable behaviors that need to be met to establish competency.

Statistical Analysis

Binary discriminant analysis identified which observable behaviors from the checklist (ie, independent variables) discriminated between the global assessment ratings (ie, dependent variable). T-scores were generated, with a higher t-score (>0) for an observed behavior signifying a higher probability that performance of that behavior predicted the assigned global assessment rating. A negative t-score signified that the behavior predicted a different global assessment rating, while a t-score of zero indicated that the behavior had no contribution in discriminating between global assessment ratings. In other words, when assessing students who received a global assessment rating of competent, a behavior with a t-score >0 would be highly predictive of a rating as competent. Alternatively, a behavior with a t-score <0 would signify that that the behavior was more predictive of either a borderline or not-competent rating while a t-score of zero would signify that the behavior was not predictive of any rating. Sensitivity analysis was also conducted by dropping the behaviors which had little or no contribution in predictability.[22] Analyses were performed using the "binda" package in R.[21,22]

Reliability between reviewers for use of the observable behavior checklist was examined as intraclass correlation coefficients and for categorical variables using Kappa statistics, with analyses performed in SAS 9.4 (SAS Institute, Cary, NC). Through our overall strategy for checklist generation, reviewer selection, and establishing reviewer reliability, we strove to establish content and internal structure validity of our assessment approach.

Results

Demographics

Twenty-six students elected to participate. Most participants reported ages between 25 and 29 (N = 23, 88%), and skewed towards female (N = 16, 62%). Students self-identified as Caucasian (77%), Asian (11%), mixed (8%), or Hispanic (4%).

Standard Setting

For the global assessment of performance, 14 students were rated as competent, 9 as borderline, and 3 as not competent. None of the "reporter" findings on the checklist were predictive of performance. The binary discriminant analysis examining the eight observable behaviors representing the "interpreter" findings is presented in Figure 2. Correctly stating that the patient required an escalation of care (highest t-score) had the largest contribution toward predicting that the student would be rated as competent. In addition, correct interpretation of vital

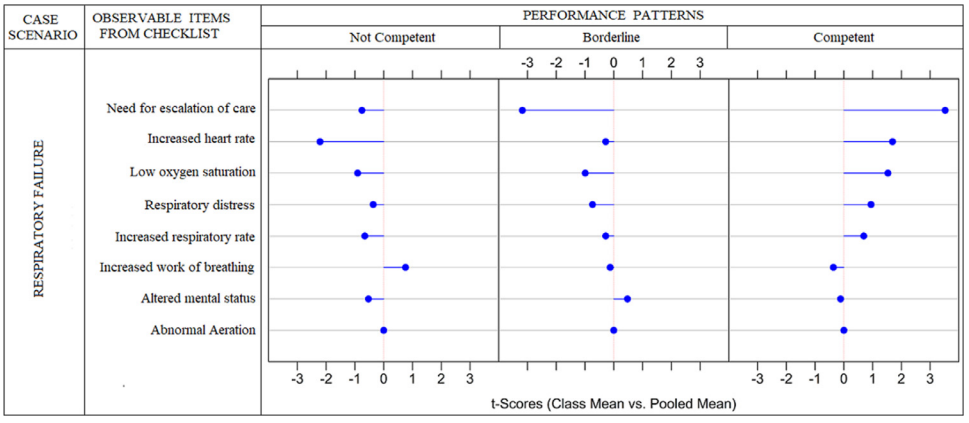

Figure 2. *Degree of predictability of observable behaviors for the global assessment of student performance. The distance away from the midline (t-score) for each observed behavior corresponds to the degree that behavior predicts the global assessment of performance. Observable behaviors at or near the midline had minimal to no contribution to the prediction of the global assessment of competence.*

signs (ie, increased heart rate, low oxygen saturation, and increased respiratory rate) and stating that the patient was in respiratory distress were other factors that predicted a rating of competent (competent performance pattern). T-scores <0 for these 5 factors differentiated students into either the borderline or not-competent categories. The addition of a positive t-score for recognition of increased work of breathing predicted a rating of not-competent category (not-competent performance pattern) while a positive t-score for recognizing altered mental status predicted a rating of borderline (borderline performance pattern). The abnormal aeration with a zero t-score (vertical midline in each performance) had no contribution in predicting the global assessment.

The degree to which student performance of these 8 observable behaviors predicted the global assessment of performance ratings is seen in Table. The predicted probability that a student will be assigned into the not-competent category based on exhibiting the not-competent performance pattern of observable behaviors was 74%. For the borderline and competent categories, the predictive probabilities were 69% and 91%, respectively for the borderline and competent performance patterns. A further sensitivity analysis, excluding the observable behaviors of recognizing increased work of breathing, altered mental status, and abnormal aeration (those findings with the smallest t-scores) yielded similar results (Appendix II and III).

Reviewer Reliability

Good reliability was demonstrated for the complete checklist of observable behaviors with an intraclass correlation coefficients of 0.71. When examining agreement between the 2 raters for each of the 8 behaviors identified through binary discriminant analysis that predicted global performance, the reliability ranged between very good agreement for recognition of respiratory distress and altered mental status (kappa = 1) to moderate agreement for increased heart rate (kappa = 0.66) (Appendix I).

Discussion

This study demonstrated the novel use of immersive VR to identify objective standards for performance assessment, moving toward the goals set forth by the AAMC Core EPAs for Entering Residency.[2] Our standard setting approach defined observable behaviors that demonstrate a high correlation with global performance ratings. Evaluators can leverage these key observable behaviors to form the basis of an objective assessment metric that may predict, and potentially replace, subjective global assessments of competency related to assessment of respiratory distress.

VR may represent a modality that can begin to close the competency assessment gap by providing a realistic environment with sufficient fidelity to prompt learners to display behaviors they would perform in a true clinical encounter. Our use of binary discriminant analysis allowed identification of the key observable behaviors that predict performance as opposed to *a priori* weighting of factors which may introduce investigator bias into assessment tool development. Our approach may serve as a strategy for medical educators to define objective measures of performance that corroborate subjective global assessments. Binary discriminant analysis can be applied in other training or assessment scenarios to identify patterns of performance that can be related to a defined category of an outcome. Such experiences could involve observed patient encounters, standardized patients, or even mannequin simulation.

Our study has several limitations. First, it was performed at a single site with 26 participants who were mainly rated as competent, limiting generalizability due to potential selection bias and applicability to less competent students. Specifically, our sample did not allow the establishment of meaningful discriminators between the borderline and not-competent groups, limiting use of our current findings for establishing passing standards. Second, while we have identified which objective behaviors predict the global assessment rating for this cohort of students, we have no evidence that global assessment ratings correlate with actual clinical performance. We have elucidated the objective findings that informed the global ratings at our institution, but these may or may not be consistent across training programs. Replication of this study across multiple institutions, generating a robust collection of student performance data and global assessment ratings, could help establish comprehensive observable behaviors that define competence for these clinical skills across programs. Third, our clinical scenario was limited to an infant with bronchiolitis. While this limits our ability to generalize to all respiratory distress, the key characteristics that define impending respiratory failure and the need for an escalation of care are consistent across underlying etiologies (eg, pneumonia, asthma, or sepsis) and patient age. Finally, VR is a resource that may not be available at all institutions. However, VR is becoming more affordable than modern computerized manikins

Table. Predicted Probabilities (95% Confidence Interval) for Receiving a Global Assessment Rating (Not Competent, Borderline, or Competent), Based on the Scores Computed for Students' Performance for Each of the Eight Observable Behavior From the Checklist, Using Binary Discriminant Analysis

	Global Assessment of Performance Rating		
Performance Rating Based on Observed Behaviors	Not Competent	Borderline	Competent
Not competent	**0.74 (0.57–0.90)**	0.24 (0.04–0.44)	0.02 (0.00–0.07)
Borderline	0.15 (0.08–0.22)	**0.69 (0.64–0.75)**	0.16 (0.09–0.22)
Competent	0.03 (0.01–0.05)	0.06 (0.04–0.08)	**0.91 (0.88–0.95)**

and standardized patients, with potentially greater opportunity for realism in illness scenarios. Additionally, VR content is easily and rapidly disseminated, and can be used remotely by learners—a functionality we now have greater appreciation for secondary to the COVID-19 pandemic.

Despite these limitations, we believe this study serves as an important early step in demonstrating the potential for VR technology to establish objective and observable competency standards for a medical student EPA. VR can overcome limitations of traditional SBME, and through this study was demonstrated as a practical method for implementing an objective assessment. Expansion of this approach may represent an effective strategy to enhance capacity for objective performance assessments − a vital step in our pursuit of CBME and ensuring students can be entrusted to provide safe and reliable care for patients upon entering residency.

ACKNOWLEDGMENTS

The authors would like to thank the medical students from the University of Cincinnati College of Medicine for their participation in this study. This study was supported in part by funding through the Council on Medical Student Education in Pediatrics (COMSEP). Funders played no role in the design and conduct of this study; collection, management, analysis, nor interpretation of the data; nor preparation, review, or approval of this article.

Funding statement: This work received project support funding from the Council on Medical Student Education in Pediatrics.

SUPPLEMENTARY DATA

Supplementary data related to this article can be found online at https://doi.org/10.1016/j.acap.2020.10.010.

REFERENCES

1. Englander R, Frank JR, Carraccio C, et al. Toward a shared language for competency-based medical education. *Med Teach*. 2017;39:582–587.
2. Englander R, Flynn T, Call S, et al. Toward defining the foundation of the MD degree: core entrustable professional activities for entering residency. *Acad Med*. 2016;91:1352–1358.
3. Powell DE, Carraccio C. Toward competency-based medical education. *N Engl J Med*. 2018;378:3–5.
4. Akaike M, Fukutomi M, Nagamune M, et al. Simulation-based medical education in clinical skills laboratory. *J Med Invest*. 2012;59:28–35.
5. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. *JAMA*. 2011;306:978–988.
6. McGaghie WC, Issenberg SB, Cohen ER, et al. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Acad Med*. 2011;86:706–711.
7. Federation of State Medical Boards of the United States I, and the National Board of Medical, Examiners. United States Medical Licensing Exam: Step 2 Clinical Skills (CS) - Content Description and General Information. 2019. Available at: https://www.usmle.org/pdfs/step-2-cs/cs-info-manual.pdf. Accessed April 13, 2020.
8. Association of American Medical Colleges. Curriculum Reports: SP/OSCE Final Examinations at US Medical Schools. 2020. Available at: https://www.aamc.org/data-reports/curriculum-reports/interactive-data/sp/osce-final-examinations-us-medical-schools. Accessed April 13, 2020.
9. Florin TA, Plint AC, Zorc JJ. Viral bronchiolitis. *Lancet*. 2017;389:211–224.
10. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Acad Med*. 2004;79(suppl 10):S70–S81.
11. Sales B, Machado L, Moraes R. Interactive collaboration for virtual reality systems related to medical education and training. *Technol Med Sci*. 2011;2011:157–162.
12. Kleinsmith A, Rivera-Gutierrez D, Finney G, et al. Understanding empathy training with virtual patients. *Comput Hum Behav*. 2015;52:151–158.
13. Real FJ, DeBlasio D, Beck AF, et al. A virtual reality curriculum for pediatric residents decreases rates of influenza vaccine refusal. *Acad Pediatr*. 2017;17:431–435.
14. Real FJ, DeBlasio D, Ollberding NJ, et al. Resident perspectives on communication training that utilizes immersive virtual reality. *Educ Health (Abingdon)*. 2017;30:228–231.
15. Zackoff MW, Real FJ, Cruse B, et al. Medical student perspectives on the use of immersive virtual reality for clinical assessment training. *Acad Pediatr*. 2019;19:849–851.
16. Zackoff MW, Real FJ, Sahay RD, et al. Impact of an immersive virtual reality curriculum on medical students' clinical assessment of infants with respiratory distress. *Pediatr Crit Care Med*. 2020;21:477–485.
17. Zackoff MW, Lin L, Israel K, et al. The future of onboarding: implementation of immersive virtual reality for nursing clinical assessment training. *J Nurses Prof Dev*. 2020;36:235–240.
18. Rothman AI, Cohen R. A comparison of empirically- and rationally-defined standards for clinical skills checklists. *Acad Med*. 1996;71(suppl 10):S1–S3.
19. Kaufman DM, Mann KV, Muijtjens AM, et al. A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Acad Med*. 2000;75:267–271.
20. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)−a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42:377–381.
21. Cox DR. The analysis of multivariate binary data. *J R Stat Soc*. 1972;C(21):113–120.
22. Gibb S, Strimmer K. Differential protein expression and peak selection in mass spectrometry data by binary discriminant analysis. *Bioinformatics*. 2015;31:3156–3162.