E-ASSESS: Creating an EPA Assessment Tool for Structured Simulated Emergency Scenarios

Caroline Andler, MD
Sneha Daya, MD
Katie Kowalek, MD
Christy Boscardin, PhD
Sandrijn M. van Schaik, MD, PhD

ABSTRACT

Background The entrustable professional activity (EPA) assessment framework allows supervisors to assign entrustment levels to physician trainees for specific activities. Limited opportunity for direct observation of trainees hampers entrustment decisions, in particular for infrequently performed activities. Simulation allows for direct observation, so tools to assess performance of EPAs in simulation could potentially provide additional data to complement clinical assessments.

Objective We developed and collected validity evidence for a simulation-based tool grounded in the EPA framework.

Methods We developed E-ASSESS (EPA Assessment for Structured Simulated Emergency ScenarioS) to assess performance in 2 EPAs among pediatric residents participating in simulation-based team training in 2017–2018. We collected validity data, applying Messick's unitary view. Three raters used E-ASSESS to assign entrustment levels based on performance in simulation. We compared those ratings to entrustment levels assigned by clinical supervisors (different from the study raters) for the same residents on a separate tool designed for clinical practice. We calculated intraclass correlation (ICC) for each tool and Pearson correlation coefficients to compare ratings between tools.

Results Twenty-eight residents participated in the study. The ICC between the 3 raters for entrustment ratings on E-ASSESS ranged from 0.65 to 0.77, while ICC among raters of the clinical tool were 0.59 and 0.57. We found no significant correlations between E-ASSESS ratings and clinical practice ratings for either EPA ($r = -0.35$ and 0.38, $P > .05$).

Conclusions Assessment following an EPA framework in the simulation context may be useful to provide data points to inform entrustment decisions as part of resident assessment.

Introduction

Entrustable professional activities (EPAs) are gaining popularity as a framework for competency-based assessment in medical education. EPAs, "units of professional practice that constitute what clinicians do as daily work,"[1] help supervisors assess trainee competency by determining how much they entrust a trainee to perform a specific activity independently. EPAs operationalize competencies by focusing on activities and associated tasks that can be observed in specific clinical contexts.[1,2] Specialty-specific EPAs have been developed for graduate medical education in several fields, including pediatrics, obstetrics and gynecology, surgery, psychiatry, internal medicine, and family medicine.[3]

One challenge with clinical performance assessments is that opportunities for direct observation in the clinical setting are declining[4]; therefore, a supervisor might be asked to make entrustment decisions without sufficient observation of a trainee's performance in a particular EPA. Simulation-based education is frequently used to augment clinical learning experiences and allow for direct observation and assessment.[5,6] Numerous tools exist for skill assessment in simulation.[7] These tend to be focused on technical or non-technical skills with checklists to identify whether the learner performed certain steps, rather than informed decisions about a learner's readiness for independent practice. It has been suggested that simulation can be used to inform entrustment decisions around specific EPAs, but this is controversial and largely untested.[8,9] To our knowledge, no published assessment tool for use in simulation has applied the EPA framework to align incidental performance evaluations in simulation with longitudinal evaluation data from clinical contexts. If we can gain reliable information about trainees' performance of specific EPAs in simulations, this may provide additional data points to make entrustment decisions. We therefore developed the E-ASSESS (EPA Assessment for Structured Simulated Emergency ScenarioS) tool, and collected validity evidence to support the use of simulation to provide assessment

DOI: http://dx.doi.org/10.4300/JGME-D-19-00533.1

Editor's Note: The online version of this article contains the E-ASSESS (EPA Assessment for Structured Simulated Emergency ScenarioS) tool.

information that can potentially contribute to entrustment decisions.

What was known and gap
The entrustable professional activity (EPA) assessment framework allows supervisors to assign entrustment levels to trainees for specific activities, but there are few opportunities for direct observation of trainees.

What is new
A simulation-based tool grounded in the EPA framework.

Limitations
Study conducted at a single institution, limiting generalizability. Only 2 EPAs were studied; better alignment might exist with other EPAs.

Bottom line
The E-ASSESS tool was easy to use and had reasonable interrater reliability, but there was no clear correlation with performance ratings for the same EPAs in clinical practice.

Methods

Setting and Participants

We conducted this project in the pediatric residency program at the University of California, San Francisco (UCSF). In July 2017, this program introduced American Board of Pediatric (ABP) EPA-based assessments for clinical supervisors to assign entrustment levels to residents they worked with during clinical rotations.[10] We modeled our E-ASSESS tool after the residency's EPA clinical practice assessment tools[11] and pilot tested it among residents who participated as leaders in an interprofessional simulation-based team training program at our institution, described in detail in a prior publication.[12] The program's learning objectives include management of acutely deteriorating patients, application of resuscitation algorithms, and effective teamwork and leadership during emergency situations. Simulation scenarios reflect common pediatric emergencies: seizure/status epilepticus, anaphylaxis, shock (hypovolemic, hemorrhagic, septic), cardiac arrest (pulseless electrical activity or arrhythmia), and respiratory failure (bronchiolitis, pneumonia, asthma exacerbation, respiratory depression).

We recruited 3 pediatricians at our institution with relevant content expertise as raters to assist with the E-ASSESS pilot. In the first phase of our study, raters reviewed video-recorded performances of a previous cohort of residents participating as team leaders in the simulation program. In the second phase, we prospectively recruited residents who participated as team leaders during the 2017–2018 academic year, video-recorded their performances for review by the study raters, and accessed their clinical practice EPA assessments provided by clinical supervisors (different from study raters).

Instrument Development

We reviewed the ABP EPAs and chose 2 applicable to activities covered in our simulation program: EPA 4, "Manage patients with acute, common diagnoses," and EPA 15, "Lead an interprofessional health care team."[10] We modeled the E-ASSESS tool (provided as online supplemental material) after our residency program's workplace-based EPA assessment tools.[11] The latter were developed by our residency leadership and use frequency-anchored questions regarding tasks and behaviors essential to the ABP EPAs and a supervision scale adapted from Chen et al.[13] E-ASSESS uses the same structure as the residency workplace-based EPA tools and consists of 3 parts: (1) behavioral items to assess specific skills integral to each EPA; (2) an entrustment scale; and (3) a free response item for the assessor to explain their reasoning. In the simulated setting a longitudinal relationship between rater and trainee is uncommon; therefore, we replaced the frequency ratings on the first part of the tool with behavioral anchors based on associated milestones.

Procedures to Collect Validity Evidence

We applied the validity framework described by Messick to our collection of validity data[14,15] and focused on 4 sources: (1) content validity; (2) response process; (3) internal structure; and (4) relationship to other variables.

Content Validity: In addition to mapping the instrument to the ABP EPAs, we developed E-ASSESS through an iterative process involving review by experts in medical education and simulation at our institution. These included pediatric subspecialists in hospital medicine, intensive care, and emergency medicine, as well as educators with PhD and Master's degrees.

Response Process: At the beginning of the study, the principal investigator (C.A.) briefed the raters on the intended use of E-ASSESS. Next, the 3 raters watched 5 video-recorded simulation scenarios and used E-ASSESS to evaluate each scenario's resident team leader. The principal investigator met with the raters and reviewed the videos using a "think-aloud protocol"[16] to explore reasons for discrepancies in ratings. We subsequently refined E-ASSESS, and the raters used the revised tool to assess resident performance in an additional 5 videos. We repeated this process for a total of 3 rounds, using different video-recorded scenarios with different resident leaders for each round.

TABLE
Intraclass Correlation for EPA Assessment Ratings Using E-ASSESS

EPA 4	Phase 1	Phase 2	EPA 15	Phase 1	Phase 2
Entrustment	0.65	0.71	Entrustment	0.67	0.77
Behavior A	0.68	0.66	Behavior A	0.45	0.66
Behavior B	0.69	0.65	Behavior B	0.50	0.65
Behavior C	0.62	0.63	Behavior C	0.75	0.82
Behavior D	0.57	0.75	Behavior D	0.62	0.69

Abbreviations: EPA, entrustable professional activity; E-ASSESS, EPA Assessment for Structured Simulated Emergency ScenarioS.
Note: Intraclass correlation coefficients (ICCs) for EPA assessment ratings for overall entrustment and specific behaviors for each of the 2 EPAs were assessed using the E-ASSESS tool. Two different cohorts of residents were included in each of the 2 study phases: phase 1 included 15 residents and phase 2 included 13 residents. Commonly cited cut-offs for qualitative ratings of agreement based on ICC: $<$ 0.40 poor, 0.40–0.59 fair, 0.60–0.74 good, and $\geq$ 0.75 excellent agreement.[17]

Internal Structure: The E-ASSESS entrustment scale allows raters to score trainees on a scale from 0 to 8, with each level correlated to an increasing level of trust in a trainee's ability to perform autonomously (from 0, trust the trainee to observe only, to 8, trust the trainee to supervise others; additional information provided as online supplemental material). We used intraclass correlation (ICC) to examine interrater reliability between the 3 raters who completed the E-ASSESS tool in both study phases.[18,19]

Relationship to Other Variables: In the second study phase, we compared ratings on E-ASSESS with entrustment ratings given by clinical supervisors on the clinical practice EPA assessment instruments for EPAs 4 and 15 during the same time frame (January–June 2018). Clinical supervisors participated in 20-minute faculty development sessions on EPAs provided by residency leadership in the year prior to the study. As the number of raters for the clinical practice tool varied for each resident and each EPA, we examined interrater reliability between these raters with 2-way random effects model ICC. We calculated mean entrustment scores for each resident across all raters for each tool and each EPA. We used Spearman's rank correlation coefficient to examine the relationship between the 2 sets of data separately for each EPA. We used SPSS Statistics 26 (IBM Corp, Armonk, NY) for all statistical analyses.

The UCSF Institutional Review Board approved the study.

Results

A total of 28 residents participated in the study: 15 in the first study phase and 13 in the second. In the second phase, 3 residents participated as simulation team leaders twice, for a total of 16 video-recorded performances in this phase. The number of ratings per resident from supervisors in the clinical setting ranged from 0 to 8 for each EPA. Two residents received no clinical practice ratings on EPA 15, and 2 residents had no ratings for EPA 4.

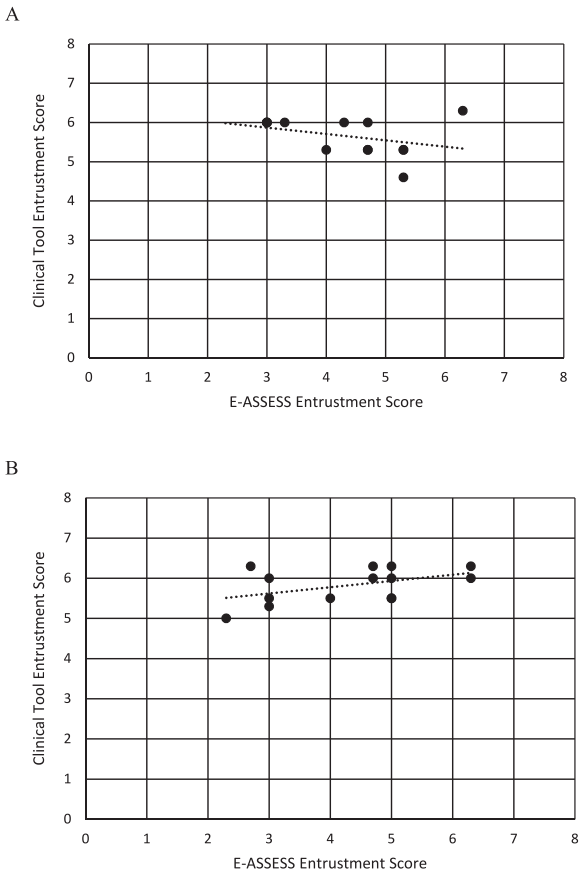
The TABLE summarizes the E-ASSESS ICC. Using commonly cited cut-offs,[20] overall agreement between the 3 raters on E-ASSESS was good for all entrustment levels. For specific behaviors within each EPA, the agreement ranged from fair to excellent. ICC among raters of the clinical practice tool was fair for both EPA 4 and EPA 15 (0.59 and 0.57, respectively).

The FIGURE shows entrustment levels on E-ASSESS versus the clinical practice instruments. The correlations between E-ASSESS ratings and clinical practice ratings were not statistically significant ($r = -0.35$ and 0.38, $P > .05$ for both correlations).

Discussion

Our E-ASSESS tool, developed to assess resident performance of 2 EPAs during simulation, appeared easy to use and had reasonable interrater reliability, but we did not find significant correlations between ratings on E-ASSESS and clinical practice assessment tools. This finding has several potential explanations worth exploring. It is possible that either E-ASSESS or the clinical practice tool (or both) do not provide a reliable assessment of the underlying constructs, at least not in the contexts in which they were used, or in the hands of the raters who used the tools. Based on the ICC data, E-ASSESS had reasonable interrater reliability, but this was less evident for the clinical practice tool. Reliability may have been compromised because ratings on the clinical practice tool may not have been based on actual observation.

In addition, unlike the raters who used E-ASSESS, raters using the clinical practice tool received limited training. Despite training, even among simulation raters, the ICC for some of the specific behaviors remained fair at best. These persistent differences in opinions among raters were likely due to their differences in professional background and expertise, which led to different expectations from learners, highlighting that rater agreement is dependent on



FIGURE

Correlations Between Entrustment Levels on E-ASSESS and Clinical Practice Tools

Note: Correlation between entrustment levels assigned on the E-ASSESS tool and the clinical practice tool for EPA 4 (panel A) and EPA 15 (panel B). Scale 0–8 for both instruments; Spearman's rho -0.35, $P = .25$ for EPA 4 and 0.38, $P = .18$ for EPA 15.

rater characteristics.[21] Of note, entrustment ratings on the clinical tool were on average much higher than ratings assigned to the same residents using E-ASSESS in simulation. This may be due to leniency bias, the phenomenon of supervisors giving overly positive assessments, typically to avoid difficult conversations or out of fear of retribution.[17,22,23] Clinical supervisors knew their evaluations would be viewed by the residents and therefore may have been prone to leniency bias, whereas study raters of simulations were told that residents would not see the ratings as they were generated for study purposes only.

A second explanation for the lack of correlation may be that E-ASSESS does not measure the same constructs as the clinical practice tool. Although both tools aim to assess the same EPAs, differences between the simulation and clinical context may lead to varying tasks and behaviors that can be observed. In most simulated scenarios, there are clear learning objectives, and the focus tends to be on the application of algorithms and/or team leadership skills within the crisis resource management framework. Real-life emergency scenarios have greater variability and are unpredictable—what is expected from team members may vary. In addition, teamwork and team leadership in clinical practice do not always center on emergencies and more often take place in low-acuity settings. While there is overlap between teamwork and team leadership skills in low- and high-acuity settings, they are not the same.[24] Considering the stakes, clinical supervisors may more easily entrust a resident with leading a team in a low-acuity setting, which is an alternative explanation for the higher ratings on the clinical practice tool for EPA 15. However, a different study in the context of our pediatric residency program found similar high ratings of leadership skills in low-acuity settings, suggesting that leniency bias may be important.[25] It is also possible that raters in the simulated setting focused on different aspects of performance than clinical supervisors, which was found to be a major contributor to interrater variability in a study examining assessment of clinical performance.[26] Rater viewpoint as well as context play an important role in how raters assess learner performance. The complexity of the clinical environment with a broad variety of sociocultural factors influencing both rater and learner performance may not lend itself well to the psychometric-based, reductionist approach of a rating scale.[27] This further complicates comparison between performance in clinical and simulation contexts.

Our study's limitations include the single institution origin, with a small sample of pediatric residents, which limited the power as well as the generalizability of our study. We also only examined 2 EPAs: other EPAs may show better alignment between simulation and clinical practice. Lastly, the residency administration provided us with clinical practice EPA ratings in a fashion that did not disclose the raters' identity; thus, it is possible that some residents received ratings from the same supervisor.

While there is some evidence that performance of procedural skills in the simulated setting may translate to real patient care settings,[6] this is less clear for other competency domains.[28] Whether simulation can be used to inform entrustment decisions is therefore controversial. The strength of simulation is that it allows for structured scenarios, a controlled environment, and limited variability, facilitating both rater training and benchmarking. Performance in one simulated scenario does not necessarily predict performance in other scenarios, and certainly not in the complexity of clinical practice. Thus, serial

assessments in multiple contexts are likely needed to inform entrustment decisions in a program approach to resident assessment.[29] Such an approach relies on multiple data points. If additional studies provide validity evidence, E-ASSESS and similar tools may be useful adjuncts to clinical practice assessments.[30] The number of data points needed to predict future performance and the relative weight one can give simulation-based assessments will require further study.

Conclusions

In this study, the E-ASSESS tool used to assess pediatric residents' performance in 2 EPAs in a simulation setting was easy to use and had reasonable interrater reliability, although there was no clear correlation with performance ratings for the same EPAs in clinical practice. The E-ASSESS tool may be a model for other similar tools to inform entrustment decisions about resident readiness for independent practice.

References

1. ten Cate O, Hoff RG. From case-based to entrustment-based discussions. *Clin Teach*. 2017;14(6):1–5. doi:10.1111/tct.12710.
2. ten Cate O. Entrustability of professional activities and competency-based training. *Med Educ*. 2005;39(12):1176–1177. doi:10.1111/j.1365-2929.2005.02341.x.
3. Kwan J, Crampton R, Mogensen LL, Weaver R, van der Vleuten CP, Hu WC. Bridging the gap: a five stage approach for developing specialty-specific entrustable professional activities. *BMC Med Educ*. 2016;16:117. doi:10.1186/s12909-016-0637-x.
4. Fromme HB, Karani R, Downing SM. Direct observation in medical education: review of the literature and evidence for validity. *Mt Sinai J Med*. 2009;76(4):365–371. doi:10.1002/msj.20123.
5. Motola I, Devine LA, Chung HS, Sullivan JE, Issenberg SB. Simulation in healthcare education: a best evidence practical guide. AMEE Guide No. 82. *Med Teach*. 2013;35(10):1511–1530. doi:10.3109/0142159X.2013.818632.
6. Brydges R, Hatala R, Zendejas B. Linking simulation-based educational assessments and patient-related outcomes: a systematic review and meta-analysis. *Acad Med*. 2015;90(2):246–256. doi:10.1097/ACM.0000000000000549.
7. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. *Acad Med*. 2013;88(6):872–883. doi:10.1097/ACM.0b013e31828ffdcf.
8. Cianciolo AT, Kegg JA. Behavioral specification of the entrustment process. *J Grad Med Educ*. 2013;5(1):10–12. doi:10.4300/JGME-D-12-00158.1.
9. Tiyyagura G, Balmer D, Chaudoin L, Kessler D, Khanna K, Srivastava G, et al. The greater good: how supervising physicians make entrustment decisions in the pediatric emergency department. *Acad Pediatr*. 2014;14(6):597–602. doi:10.1016/j.acap.2014.06.001.
10. The American Board of Pediatrics. Entrustable Professional Activities for General Pediatrics. https://www.abp.org/sites/abp/files/pdf/combined_gp_epas.pdf. Accessed February 26, 2020.
11. West DC, Henry D, Laves E, McNamara M. Development and implementation of a workplace-based assessment system to inform competency decisions and encourage self-regulated and mastery learning behaviors in post-graduate medical education. 2019 Association for Medical Education in Europe (AMEE) Meeting. Abstract 1742. https://amee.org/getattachment/Conferences/AMEE-2019/Abstracts/AMEE-2019-Abstract-Book-Post-Conference-v2.pdf. Accessed February 26, 2020.
12. van Schaik SM, Plant J, Diane S, Tsang L, O'Sullivan P. Interprofessional team training in pediatric resuscitation: a low-cost, in situ simulation program that enhances self-efficacy among participants. *Clin Pediatr (Phila)*. 2011;50(9):807–815. doi:10.1177/0009922811405518.
13. Chen HC, van den Broek WES, ten Cate O. The case for use of entrustable professional activities in undergraduate medical education. *Acad Med*. 2015;90(4):431–436. doi:10.1097/ACM.0000000000000586.
14. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ*. 2003;37(9):830–837. doi:10.1046/j.1365-2923.2003.01594.x.
15. Messick S. Standards of validity and the validity of standards in performance assessment. *Educ Measure Issues Prac*.1995;14(4):5–8.
16. Charters E. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Educ*. 2003;12(2):68–82.
17. Bandiera G, Lendrum D. Daily encounter cards facilitate competency-based feedback while leniency bias persists. *CJEM*. 2008;10(1):44–50. doi:10.1017/s1481803500010009.
18. Boscardin CK, Wijnen-Meijer M, ten Cate O. Taking rater exposure to trainees into account when explaining rater variability. *J Grad Med Educ*. 2016;8(5):726–730. doi:10.4300/JGME-D-16-00122.1.
19. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor*

Quant Methods Psychol. 2012;8(1):23–34. doi:10.20982/tqmp.08.1.p023.
20. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6(4):284–290.
21. Lumley T, McNamara TF. Rater characteristics and rater bias: implications for training. *Lang Test*. 1995;12(1):54–71. doi:10.1177/026553229501200104.
22. van Schaik SM, Boscardin CK, O'Brien B, Adler S. Interprofessional teamwork: are we ready for skills assessment? *OSF Preprints*. https://osf.io/mfqs2. Accessed February 26, 2020.
23. Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. *Acad Med*. 2005;80(10 suppl):84–87. doi:10.1097/00001888-200510001-00023.
24. van Schaik SM, O'Brien BC, Almeida SA. Perceptions of interprofessional teamwork in low-acuity settings: a qualitative analysis. *Med Educ*. 2014;48(6):583–592. doi:10.1111/medu.12424.
25. Oza SK, van Schaik SM, Boscardin CK, Pierce R, Miao E, Lockspeiser T, et al. Leadership observation and feedback tool: a novel instrument for assessment of clinical leadership skills. *J Grad Med Educ*. 2018;10(5):573–582. doi:10.4300/JGME-D-18-00113.1.
26. Gingerich A, Ramlo SE, van der Vleuten CPM, Eva KW, Regehr G. Inter-rater variability as mutual disagreement: identifying raters' divergent points of view. *Adv Health Sci Educ Theory Pract*. 2017;22(4):819–838. doi:10.1007/s10459-016-9711-8.
27. Govaerts M, van der Vleuten CPM. Validity in work-based assessment: expanding our horizons. *Med Educ*. 2013;47(12):1164–1174. doi:10.1111/medu.12289.
28. Tavares W, Brydges R, Myre P, et al. Applying Kane's validity framework to a simulation based assessment of clinical competence. *Adv Health Sci Educ Theory Pract*. 2018;23(2):323–338. doi:10.1007/s10459-017-9800-3.
29. van der Vleuten CP. Revisiting 'assessing professional competence: from methods to programmes.' *Med Educ*. 2016;50(9):885–888. doi:10.1111/medu.12632.
30. Moonen–van Loon JM, Overeem K, Govaerts MJ. The reliability of multisource feedback in competency-based assessment programs: the effects of multiple occasions and assessor groups. *Acad Med*. 2015;90(8):1093–1099. doi:10.1097/ACM.0000000000000763.

Caroline Andler, MD, is Clinical Fellow, Pediatric Emergency Medicine, Children's Hospital Los Angeles; **Sneha Daya, MD,** is Assistant Professor of Internal Medicine and Pediatrics, The University of California, San Francisco (UCSF); **Katie Kowalek, MD,** is Assistant Professor of Pediatrics, Division of Pediatric Critical Care Medicine, University of Arizona; **Christy Boscardin, PhD,** is Associate Professor, Department of Medicine, and Education Scientist, Center for Faculty Educators, UCSF; and **Sandrijn M. van Schaik, MD, PhD,** is Professor, Department of Pediatrics, UCSF, and Baum Presidential Chair for Experiential Learning and Director, UCSF Kanbar Center for Simulation and Clinical Skills.

Funding: The authors report no external funding source for this study.

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

The authors would like to thank the staff of the University of California, San Francisco Kanbar Simulation Center, faculty, and participants in the University of California, San Francisco Health Professions Education Pathway Program, and all the residents who participated in the study.

Corresponding author: Sandrijn M. van Schaik, MD, PhD, University of California, San Francisco, 550 16th Street, Floor 5, Box 0106, San Francisco, CA 94143, 415.476.3731, sandrijn.vanschaik@ucsf.edu

Received July 30, 2019; revisions received December 2, 2019, and January 28, 2020; accepted January 31, 2020.