The Reliability of Rater Variability

Andrea Gingerich, PhD

Simulation is well recognized for its affordances for collecting important assessment information.[1–3] In this issue of the *Journal of Graduate Medical Education*, Andler and colleagues present validity evidence for leveraging the simulation context to provide assessment data for entrustable professional activities (EPAs).[4] Unfortunately, they found their validity argument hampered by an unexpected finding: despite good interrater reliability for entrustment-based simulation assessment ratings and fair interrater reliability for similar entrustment-based clinical practice ratings, there were no correlations between them. The authors ponder possible explanations for this troublesome finding and suggest that since there was only "fair agreement at best" for some of the behaviors, rater variability might be an explanation for the lack of correlations.

The havoc that rater variability has inflicted on reliability measures has spurred several of us to study its sources.[5–7] Aspects not directly related to the rating scale, such as the context in which assessments take place[8–10] and variations in rater interpretations and judgments,[11–14] have been identified as contributors to rater variability. Thus, I am not surprised to see rater variability when an entrustment scale is used. In fact, as evidence of rater variability continues to accumulate along with increasing recognition of the "plurality of interpretations,"[15] we may be reaching a point where rater variability can no longer be framed as an unexpected finding. Yet, this raises a conundrum for the assessment field. Accepting rater variability as the status quo would complicate plans for collecting and interpreting validity evidence.[16] How can we demonstrate a relationship to other variables without reliability?

In part, the simulation context might offer a solution to this by providing a stable context where raters can be standardized and, themselves, judged. Almost 2 decades ago,[17] medical educators were directed to techniques that optimize interrater reliability—figure skating judging.[18,19] Although it is not free from bias,[20] figure skating judging has design features that support rater agreement and interrater reliability. First, judges are trained and monitored so that those who share consensus are invited to continue judging and outlier judges are not. Second, the assessed performance lasts only a few minutes with a specified number of predictable elements that can be performed in a limited number of ways, with each variation assigned a corresponding score. Third, the assessment task is the judge's only task where they directly observe a series of similar performances. They assign ratings immediately after each assessment, and then note how their ratings compare with those of other judges. These design features are incompatible with almost every aspect of workplace-based assessment; however, the simulation context does offer similar affordances.[21] Yet, I wonder how the design features that aim to minimize all types of unwanted variability would align with the very notion of entrustment-based assessment?

Entrustment, entrustability, and level of supervision scales promised to better mimic the judgments and decisions supervisors make in the workplace.[22,23] The construct of entrustment resonated with the essence of supervision.[24,25] It offered to systematically track subjective expert judgments of overall performance to complement the competence judgments based on observed behaviors that were already being collected and analyzed.[26] I was excited about using entrustment as the basis for workplace-based assessment because it had the potential to capture indescribable and nuanced aspects of being a physician that resisted measurement.[27] I am not an expert in simulation so I will pose the question to those who are: How well does entrustment align with what raters are doing, thinking, and feeling during simulation? It is not a straightforward question and leads to other difficult questions. What does it mean to entrust in simulation and how does it compare to entrusting in the workplace? For example, is the construct of entrustment most aligned when the rater is exposed to the competing priorities of patient safety, learner autonomy, clinical care, teaching obligations, service efficiency, and learner welfare? In other words, must the rater be simultaneously engaged with supervising the trainee for the construct of entrustment to be sufficiently aligned? If so, which forms of simulation offer that context for raters?

In proposing that entrustment can be used as the basis for assessment in simulation, the latest research of Andler and colleagues offers the opportunity to contemplate the ideal constructs for simulation

DOI: http://dx.doi.org/10.4300/JGME-D-20-00163.1

assessment. If we were without contemporary pressures to provide data to inform EPA decisions, would we choose to use entrustment in this context? The assessment construct of feedback provision (like that used by field notes[28]) may be better aligned than entrustment if the rater's role in simulation is akin to that of a coach helping a trainee to learn during practice. Or perhaps the predictable and controllable conditions of simulation, similar to that of figure skating judging, could be used to optimize measurement of competence through standardized assessment of performance.

Entrustment-based assessment is rapidly becoming an important component of our assessment tool kit, but I cannot imagine a post-psychometric utopia where all assessments are based on entrustment. All of our assessment modalities (including EPAs), assessment constructs (including entrustment), and assessment contexts (including simulation) have strengths to be leveraged and limitations to be accommodated. Fortunately, the limitations of one can be strategically addressed by the strengths of another with its own limitations supported by yet another context or construct or modality.[29] I am eager to see how the strengths of the simulation assessment context and the construct of entrustment can contribute to an assessment program that is more informative than the sum of its parts.

References

1. Amin Z, Boulet JR, Cook DA, Ellaway R, Fahal A, Kneebone R, et al. Technology-enabled assessment of health professions education: consensus statement and recommendations from the Ottawa 2010 conference. *Med Teach*. 2011;33(5):364–369. doi:10.3109/0142159X.2011.565832.
2. St-Onge C, Lineberry M. Simulation for assessment. In: Chiniara G, ed. *Clinical Simulation*. 2nd ed. Cambridge, MA: Academic Press; 2019:867–877.
3. Boulet JR. Summative assessment in medicine: the promise of simulation for high-stakes evaluation. *Acad Emerg Med*. 2008;15(11):1017–1024. doi:10.1111/j.1553-2712.2008.00228.x.
4. Andler C, Kowalek K, Boscardin C, van Schaik SM. E-ASSESS: creating an EPA assessment tool for structured simulated emergency scenarios. *J Grad Med Educ*. 2020;12(2):153–158.
5. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box' differently: assessor cognition from three research perspectives. *Med Educ*. 2014;48(11):1055–1068. doi:10.1111/medu.12546.
6. Gauthier G, St-Onge C, Tavares W. Rater cognition: review and integration of research findings. *Med Educ*. 2016;50(5):511–522. doi:10.1111/medu.12973.
7. Lee V, Brain K, Martin J. Factors influencing Mini-CEX rater judgments and their practical implications: a systematic literature review. *Acad Med*. 2017;92(6):880–887. doi:10.1097/ACM.0000000000001537.
8. Gingerich A. Comparatively salient: examining the influence of preceding performances on assessors' focus and interpretations in written assessment comments. *Adv Health Sci Educ Theory Pract*. 2018;23(5):937–959. doi:10.1007/s10459-018-9841-2.
9. Lee V, Brain K, Martin J. From opening the 'black box' to looking behind the curtain: cognition and context in assessor-based judgements. *Adv Health Sci Educ Theory Pract*. 2019;24(1):85–102. doi:10.1007/s10459-018-9851-0.
10. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model. *Med Educ*. 2011;45(10):1048–1060. doi:10.1111/j.1365-2923.2011.04025.x.
11. Govaerts MJ, Van de Wiel MW, Schuwirth LW, Van der Vleuten CP, Muijtjens AM. Workplace-based assessment: raters' performance theories and constructs. *Adv Health Sci Educ Theory Pract*. 2013;18(3):375–396. doi:10.1007/s10459-012-9376-x.
12. St-Onge C, Chamberland M, Lévesque A, Varpio L. Expectations, observations, and the cognitive processes that bind them: expert assessment of examinee performance. *Adv Health Sci Educ Theory Pract*. 2016;21(3):627–642. doi:10.1007/s10459-015-9656-3.
13. Tavares W, Ginsburg S, Eva KW. Selecting and simplifying: rater performance and behavior when considering multiple competencies. *Teach Learn Med*. 2016;28(1):41–51. doi:10.1080/10401334.2015.1107489.
14. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. *Adv Health Sci Educ Theory Pract*. 2013;18(3):325–341. doi:10.1007/s10459-012-9372-1.
15. Hodwitz K, Kuper A, Brydges R. Realizing one's own subjectivity: assessors' perceptions of the influence of training on their conduct of workplace-based assessments. *Acad Med*. 2019;94(12):1970–1979. doi:10.1097/ACM.0000000000002943.
16. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. *Adv Simul (Lond)*. 2016;1:31. doi:10.1186/s41077-016-0033-y.
17. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med*. 2003;15(4):270–292. doi:10.1207/S15328015TLM1504_11.

18. Huang J, Foote CJ. Using generalizability theory to examine scoring reliability and variability of judging panels in skating competitions. *J Quant Analy Sport*. 2011;7(3). https://doi.org/10.2202/1559-0410.1241. Accessed February 24, 2020.
19. Weekley JA, Gier JA. Ceiling in the reliability and validity of performance ratings: the case of expert raters. *Acad Manag J*. 1989;32(1):213–222. doi:10.5465/256428.
20. Sala BR, Scott JT, Spriggs JF. The Cold War on ice: constructivism and the politics of Olympic figure skating judging. *Perspect Politics*. 2007;5(1):17–29. doi:10.1017/S153759270707003X.
21. Weersink K, Hall AK, Rich J, Szulewski A, Dagnone JD. Simulation versus real-world performance: a direct comparison of emergency medicine resident resuscitation entrustment scoring. *Adv Simul (Lond)*. 2019;4(1):9. doi:10.1186/s41077-019-0099-4.
22. ten Cate O. Nuts and bolts of entrustable professional activities. *J Grad Med Educ*. 2013;5(1):157–158. doi:10.4300/JGME-D-12-00380.1.
23. Rekman J, Gofton W, Dudek N, Gofton T, Hamstra SJ. Entrustability scales: outlining their usefulness for competency-based clinical assessment. *Acad Med*. 2016;91(2):186–190. doi:10.1097/ACM.0000000000001045.
24. Pangaro L, ten Cate O. Frameworks for learner assessment in medicine: AMEE guide no. 78. *Med Teach*. 2013;35(6):e1197–e1210. doi:10.3109/0142159X.2013.788789.
25. Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. *Med Educ*. 2011;45(6):560–569. doi:10.1111/j.1365-2923.2010.03913.x.
26. Ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med*. 2007;82(6):542–547. doi:10.1097/ACM.0b013e31805559c7.
27. Gingerich A. What if the 'trust' in entrustable were a social judgement? *Med Educ*. 2015;49(8):750–752. doi:10.1111/medu.12772.
28. Ross S, Poth CN, Donoff M, Humphries P, Steiner I, Schipper S, et al. Competency-based achievement system: using formative feedback to teach and assess family medicine residents' skills. *Can Fam Physician*. 2011;57(9):e323–e330.
29. Van der Vleuten C, Schuwirth L, Driessen E, Dijkstra J, Tigelaar D, Baartman LK, et al. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205–214. doi:10.3109/0142159X.2012.652239.

Andrea Gingerich, PhD, is Assistant Professor, Northern Medical Program, University of Northern British Columbia, Prince George, British Columbia, Canada.

Corresponding author: Andrea Gingerich, PhD, Northern Medical Program, 3333 University Way, Prince George, British Columbia V2N 4Z9, Canada, 250.960.5432, andrea.gingerich@unbc.ca