Reframing the O-SCORE as a Retrospective Supervision Scale Using Validity Theory

Walter Tavares, PhD
Wade Gofton, MD, MEd
Farhan Bhanji, MD, MSc(Ed)
Nancy Dudek, MD, MEd

Validity is a fundamental consideration in developing and evaluating assessment programs, processes, and tools.[1] In response to ongoing validity challenges with workplace-based assessment (WBA), Gofton et al developed the Ottawa Surgical Competency Operating Room Evaluation (O-SCORE).[2] In its development, the authors argued that the ultimate goal of postgraduate medical education is to produce trainees who are competent to practice independently, and that it may be helpful to structure assessments based on this concept. Using colloquial faculty language, they created a supervision-aligned set of *anchors* that range from 1 ("I had to do. . .") to 5 ("I did not need to be there. . ."). The O-SCORE and other WBA tools that used similar types of scales demonstrated evidence of validity and seemed to be performing better than previous rating scales.[3] Given the alignment the O-SCORE scale language had with the way faculty understood the goals of training, and the conceptual link that was being made between supervision and entrustment, Rekman et al chose to describe these scales as "entrustability scales."[3] The alignment with competency-based medical education, intuitive appeal, ease of use, and early validity evidence led the O-SCORE anchors and entrustability scales in general to become widely adopted. However, they have also become a threat to assessment validity based on potential interpretation issues, in part from naming (and framing) supervision-type scales as meaning "entrustability." Indeed, the use of entrustability scales in practice has been met with confusion, concern, and criticism.[4] Since then, ten Cate et al and others have suggested a number of refinements and corrections, including the need to disentangle "retrospective" and "prospective" assessments.[5–7]

In this Perspectives article we pick up on this distinction between retrospective and prospective assessment using assessment validity as our framework and the O-SCORE as our example. Our first aim is to reframe the O-SCORE's rating scale anchors as point-in-time retrospective assessments of the faculty's experience with a trainee, and not as prospective indicators of readiness, or as immediate claims about the level of entrustability. We also aim to elaborate on how and why retrospective and prospective distinctions serve as a meaningful correction to potential misunderstandings about the O-SCORE anchors in graduate medical education. We selected the O-SCORE anchors because they may have perpetuated some interpretation issues yet continue to be widely adopted in graduate medical education.[8–11] Using Cizek's conceptualization of validity, which separates score *meaning* from score *use*,[12,13] we begin by describing the O-SCORE and how its *meaning* can be misinterpreted, to ultimately speak to its *use*.

Score Meaning

We propose the O-SCORE scale language itself could pose a validity threat if there is confusion about the *meaning* of its anchors. In its original development, the use of colloquial faculty language suggested a construct that exists with the faculty (eg, "I had to be there") rather than the trainee, even if informed by the trainee's behaviors. In assessment contexts, this can create tension if faculty confuse rater and trainee level constructs. That is, faculty may struggle with positioning the construct in the moment, trying to resolve whether the language is about "me" (the faculty) or "them" (the trainee). The distinction is a subtle but important point in extrapolating what the scores mean and how they should be used.

What then do scores generated using the O-SCORE mean? The emphasis on "I" (as in "*I* had to do") may reflect several faculty-owned influences and interpretations of the trainee's performance in complex contexts, and what the interaction means for them (ie, the faculty). Using "I" speaks to more than asking faculty to report behaviors exhibited by trainees or even matching behaviors to predefined performance expectations. The past tense language should be taken to represent a retrospective or reflective opportunity on the part of the faculty based on their experience with the learner for the encounter being assessed. On

DOI: http://dx.doi.org/10.4300/JGME-D-21-00592.1

its own, these anchors make no claims about the entrustability of a learner, or about how that learner will perform in the future. It is simply a record of the faculty's rating of how much supervision or assistance they provided in that encounter.[14]

Suggesting the O-SCORE scale in WBA permits an inferential claim or that it *means* a characteristic of the trainee—in this case entrustability—may be less accurate than saying the observations reflect what the trainee's performance and several other contextual factors meant to the faculty member. For some, it is this subjective information that then becomes valuable in the collective.[15] Completing these supervision-type scales speak only to how faculty have "made sense" of the observed performance or their experience with the trainee. Previous research exploring rater cognition has revealed this active, subjective, and personal translation process even for the same performance.[15–17] When viewed in this way, issues like error or reliability are diminished for richness of the interaction and making use of those faculty experiences and differences—a philosophical issue with practical considerations.[18] Reformulating supervision anchor language from being only about the trainee, to being more about how faculty experience the assessment activity (ie, interaction between clinical stimuli, learner, and faculty), promotes better construct alignment for faculty and for those interpreting and using the data.

Score Use

The second way to rethink the O-SCORE rating scale language is to consider its *use*. Here, now that we have clarity on what the scores *mean*, we would ask, Can or should the O-SCORE and other supervision-type scales be used for decisions related to presumptive trust? For example, How will a trainee perform in the future? Here the descriptions of retrospective and prospective assessment scales from ten Cate et al is helpful.[6] We argued above that retrospective supervision-type scales in WBA reflect faculty experiences in the moment that shape what they did, informed by their own experiences and comfort, and what they understood about the trainee in a particular context. Our task then is to align that *meaning* with *use*. As opposed to claims about presumptive trust, retrospective scales can be used for formative purposes, where all the behavioral, historical, social, personal, and contextual features that led to the faculty's action or reflection can be excavated through "learning conversations"[19] (eg, debriefing, feedback), since these are present in the assessment activity. This would be an example of alignment between score meaning and use. When this is present, interpretation issues are corrected, and validity can be optimized.

Prospective assessments represent a different intended *use*, one that involves decisions related to the trainee's ability to assume *future* responsibilities and care activities.[6] Entrustment decisions are prospective. These types of assessments can be made using a blend of retrospective assessment and other data, and are typically categorical, rather than ordinal. Competence committees serve as examples of where prospective assessments can take place. Here the *intended use* can include, for example, progression to higher levels of responsibility, access to unsupervised activities, or graduation, something point-in-time retrospective assessments alone are unable to do. Prospective assessments are more complex, structured and enacted differently than retrospective assessments, and often (or should) include more than just individual documented observations.[6] Therefore, the use of retrospective supervision scales (like the O-SCORE) have been described as having value for prospective purposes by providing data to support decisions about entrustment,[6] but are themselves limited for that use. On their own, individual retrospective supervision scales do not provide information about the complex construct of entrustment and therefore should not be misinterpreted by faculty or trainees as trying to do so.

In summary, the O-SCORE scale and other retrospective supervision-type scales may be misinterpreted, leading to confusion and tensions in practice. While tools using the O-SCORE anchors have demonstrated strong psychometric properties, we are raising issues of interpretation, specifically related to score *meaning* and *use*, and the alignment between these as important validity considerations. Rather than describing these scales as entrustability scales, which suggest a prospective type of assessment, we support the recommendation that these types of scales are best thought of and used as retrospective scales.[6,7] We would add that the reflective and faculty-aligned language and construct support this suggestion. Inferential claims or *meaning* must therefore not be about the entrustability of learners, but rather faculty reflections on their actions based on their experience with trainees for a given context and time. Whether the same assessment can and should be used for formative and summative purposes, and what the implications are of doing so, should also be examined carefully, but they do represent 2 distinct intended uses. This calls for careful action on the part of educators to avoid unintended meaning and use.

To assist further in providing clarity, while ten Cate et al suggested that these types of scales be described as "retrospective entrustment supervision scales,"[6] our concern is that this may not go far enough. Using the word entrustment in the name of the scale may perpetuate ongoing confusion for front-line faculty

who are providing these assessments about what construct is in play and when, in the same way that calling the O-SCORE scale an entrustability scale did. Instead, in an effort to mitigate validity threats by clarifying intended meaning and use, we suggest the assessment community consider describing scales that use faculty reflections of their supervision choices or reflections as just that—retrospective supervision scales.

References

1. Kane MT. Validating the interpretations and uses of test scores. *J Educ Measur.* 2013;50(1):1–73. doi:10.1111/jedm.12000
2. Gofton W, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med*. 2012;87(10):1401–1407. doi:10.1097/ACM.0b013e3182677805
3. Rekman J, Gofton W, Dudek NL, 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
4. Melvin L, Rassos J, Stroud L, Ginsburg S. Tensions in assessment: the realities of entrustment in internal medicine. *Acad Med*. 2020;95(4):609–615. doi:10.1097/ACM.0000000000002991
5. Ten Cate O. When I say entrustability. *Med Educ*. 2020;54(2):103–104. doi:10.1111/medu.14005
6. Ten Cate O, Schwartz A, Chen HC. Assessing trainees and making entrustment decisions: on the nature and use of entrustment-supervision scales. *Acad Med*. 2020;95(11):1662–1669. doi:10.1097/ACM.0000000000003427
7. Schumacher DJ, ten Cate O, Damodaran A, et al. Clarifying essential terminology in entrustment. *Med Teach*. 2021;43(7):737–744. doi:10.1080/0142159X.2021.1924365
8. Rekman J, Hamstra SJ, Dudek NL, Wood TJ, Seabrook C, Gofton W. A new instrument for assessing resident competence in surgical clinic: the Ottawa Clinic assessment tool. *J Surg Educ*. 2016;73(4):575–582. doi:10.1016/j.jsurg.2016.02.003
9. Voduc N, Dudek NL, Parker CM, Sharma KB, Wood TJ. Development and validation of a bronchoscopy competence assessment tool in a clinical setting. *Ann Am Thorac Soc*. 2016;13(4):495–501. doi:10.1513/AnnalsATS.201508-548OC
10. Royal College of Physicians and Surgeons of Canada. Work based assessment implementation guide: formative tips for medical teaching practice. https://cbmepg.mcmaster.ca/app/uploads/2020/05/work-based-assessment-practical-implications-implementation-guide-e.pdf. Accessed December 2, 2021.
11. Cheung WJ, Wood TJ, Gofton W, Dewhirst S, Dudek NL. The Ottawa Emergency Department Shift Observation Tool (O-EDShOT): a new tool for assessing resident competence in the emergency department. *AEM Educ Train*. 2020;4(4):359–368. doi:10.1002/aet2.10419
12. Cizek GJ. Validating test score meaning and defending test score use: different aims, different methods. *Assess Educ*. 2016;23(2):212–225. doi:10.1080/0969594X.2015.1063479
13. Cizek GJ. Defining and distinguishing validity: interpretations of score meaning and justifications of test use. *Psychol Methods*. 2012;17(1):31–43. doi:10.1037/a0026975
14. Robinson TJ, Wagner N, Szulewski A, Dudek NL, Cheung WJ, Hall AK. Exploring the use of rating scales with entrustment anchors in workplace-based assessment. *Med Educ*. 2021;55(9):1047–1055. doi:10.1111/medu.14573
15. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*. 2013;35(7):564–568. doi:10.3109/0142159X.2013.789134
16. 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
17. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently. *Adv Health Sci Educ Theory Pract*. 2013;18(3):325–341. doi:10.1007/s10459-012-9372-1
18. Tavares W, Kuper A, Kulasegaram K, Whitehead C. The compatibility principle: on philosophies in the assessment of clinical competence. *Adv Health Sci Educ Theory Pract*. 2020;25(4):1003–1018. doi:10.1007/s10459-019-09939-9
19. Tavares W, Eppich W, Cheng A, et al. Learning conversations: an analysis of the theoretical roots and their manifestations of feedback and debriefing in medical education. *Acad Med*. 2020;95(7):1020–1025. doi:10.1097/ACM.0000000000002932

Walter Tavares, PhD, is Assistant Professor and Scientist, The Wilson Centre and Temerty Faculty of Medicine, University Health Network and University of Toronto, Toronto, Ontario, Canada; **Wade Gofton, MD, MEd,** is Professor, Department of Surgery, University of Ottawa, Ottawa, Ontario, Canada; **Farhan Bhanji, MD, MSc(Ed),** is Professor, Department of Pediatrics, McGill University, Montreal, Quebec, Canada, and Associate Director of Assessment Strategy Royal College of Physicians and Surgeons, Ottawa, Ontario, Canada; and **Nancy Dudek, MD, MEd,** is Professor, Department of Medicine, Division of Physical Medicine and Rehabilitation, and The Ottawa Hospital, University of Ottawa, Ottawa, Ontario, Canada.

Corresponding author: Walter Tavares, PhD, University Health Network and University of Toronto, Toronto, Ontario, Canada, walter.tavares@utoronto.ca, Twitter @WalterTava