Use of Clinical Performance Rating/Checklist by US and Canadian Medical Schools

A major category of learner assessment in US and Canadian medical schools, after the two dominant categories – institutional written exams and participation grades – is clinical performance ratings/checklists. Checklist-formatted assessments seem particularly well-suited when we have clear performance expectations for learners and want to provide behaviorally-focused feedback.

On the DR-ED listserv recently, a member wondered how reliable checklist item scores should be (as opposed to the reliability of checklist totals or averages), since we often provide item-level feedback after checklist-based assessments. This is a great way to ask the "how much reliability is enough?" question, by clearly specifying, "enough
for what purpose?" After all, we use scores from checklists and other types of assessments in many ways: to label learners (e.g. "entrustable" or not), to decide on consequences (e.g. being required to remediate), to provide feedback, to see learner group trends and gauge how effectively we are performing as educators, and so on.

Perhaps surprisingly, each of those different use cases has a different degree of reliability associated with it – literally a different numeric value. We're used to reading in journal articles for instance that a certain assessment "has" a reliability of some specific number. In fact, we can only properly say how reliable a measurement’s scores are when we've decided what we'll do with them, and the same checklist can have many different reliabilities depending on how it's being used.1 (By extension: commonplace reliability statistics that we've tended to treat as "the" reliability statistic make assumptions about how we'll use the scores that often aren't aligned with our intended uses,2 and those misapplied estimates can be dramatically off.) So, defining our intended interpretations and uses of scores comes first.

Second: David Cook and I have proposed recently3 that the most important consideration for how valid an assessment is – reliability being an important part of validity – is whether the assessment's use leads to the positive outcomes we hope it will, without overly severe negative outcomes. For instance, in sharing item-level checklist feedback with learners, we might want checklist item scores to reflect genuine areas of strength and weakness for each learner, and for learners to then use those to better focus their time during studying. If a certain checklist is reliable and accurate about learners’ strengths and weaknesses, but learners don’t use that information, instead studying largely as they would have had they not received it, then high reliability wasn’t sufficient to lead to positive outcomes. Of course, if a checklist is unreliable and learners do use it to structure their studying, they may make less optimal use of their time. So, the question of what "sufficient reliability" is – as it relates to providing feedback to learners – should be asked in terms of how differing reliability relates to positive and/or negative outcomes.

Given all of this, the common "rule of thumb" that a reliability of .70-.80 is acceptable for a test or measurement is worth challenging. Since we usually see reliability statistics about total scores, it can be surprising how unreliable individual checklist items (or any items) can be within a multi-item assessment. A ten-item checklist with an overall reliability of .75, such that the total score is fairly consistent, might still have several items with extremely low reliability. However, whether an item is "acceptably" reliable or not requires closer consideration. If our learners are being rated even a little inconsistently (say, a reliability of .85) on something highly observable such as whether they "auscultated the lungs", and there is a low-cost way to improve reliability, and doing so would meaningfully benefit learning, then let's do it! Conversely, when there aren't low-cost ways to improve our measurements, we might accept some quite unreliable items (say, a reliability of .50) if we're reasonably certain that the act of measurement is still facilitating sound learning – as might be the case if we are up front with the learners about our own inconsistency and invite them to collaborate with us in appraising that aspect of their performance.

One question to consider is, "how many different pieces of feedback can learners make use of at once?" At least one recent study suggests that, for a resident-level OSCE at their institution, the answer is "not many".4 As we consider how to design our assessment systems to be acceptably reliable, we should also be carefully designing our assessment feedback to be of an appropriate breadth, depth, and presentation format for learners to be able to process it, remember it, and apply it in future learning.
Author:
Matthew Lineberry, PhD, is the Director of Simulation Research, Assessment, and Outcomes for the University of Kansas Medical Center (KUMC) and the University of Kansas Hospital, both in Kansas City, KS. Matt is also an Assistant Professor in the Department of Health Policy and Management at KUMC.

References