A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment

Reference:
Ilgen JS¹, Ma IWY², Hatala R³, Cook DA⁴-⁵. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. Medical Education, Feb 2015; 49(2): 161-73

Author Institutions:
¹Division of Emergency Medicine, Department of Medicine, University of Washington School of Medicine, Seattle, WA, USA
²Department of Medicine, University of Calgary, Calgary, AB
³Department of Medicine, University of British Columbia, Vancouver, BC
⁴Mayo Clinic Multidisciplinary Simulation Center, Mayo Clinic College of Medicine, Rochester, MN, USA
⁵Division of General Internal Medicine, Mayo Clinic, Rochester, MN, USA

Tags
Clinical domain
General

Educational domain
Assessment
Undergraduate
(medical school)
(Post)graduate
(Residency training)

Background

There is an ongoing debate in the HPE literature that is not explicit, but in my opinion, has widely infiltrated discussions involving assessment of learners. Put one way, should we standardize our approach to assessment of performance, which improves reliability but leads to issues of excessive reductionism and attention to superficial features? Or, should we trust the gestalt judgment of “experts,” which attends to the origins of medical education but leads to issues of rater bias. (For more details on the challenges of rater cognition, see KeyLIME episode 78.)

Checklists (CL) provide a scaffold for uninitiated assessors. Global rating scales (GRS) allow experts to use a more natural reasoning process (e.g., the complexity of diagnosis is neither dichotomous, nor algorithmic). Research evidence supports both sides of the debate. Depending upon your bias, you can argue that a CL forces a
systematic, and not idiosyncratic, assessment framework. In contrast, GRS are more discriminating at higher levels of expertise, presumably because they allow the input of more complex data points that cannot be easily captured in yes/no checkboxes.

This paper attempts to weigh the evidence and see if there is a unifying story.

**Purpose**

"1. What are the inter-rater, inter-item and inter-station reliabilities of global ratings in comparison with checklist scores?

2. How well do global ratings and checklist scores correlate?

3. What validity evidence has been reported for global ratings and checklist scores?

The prevailing view has held that the GRS offers greater reliability than the checklist. In the present review, we sought evidence to confirm or refute this opinion."

**Type of paper**

Systematic review
Meta-analysis

**Key Points on the Methods**

This study employed exemplary methods that adhere to the PRISMA (preferred reporting items for systematic reviews and meta-analyses) guidelines. (For a further description of the study protocol see KeyLIME Ep#44.)

The "general" inclusion criterion was simulation-based (but not standardized patients) health professions research comparing checklist to GRS. The studies included in this paper are part of a larger research program on technology-enhance simulation. The search strategy was updated to February, 2013.

**Potential CoI: JS is a co-author on two publications that used the data set discussed in this paper.**

**Key Outcomes**

- n= 45 studies; 1819 learners (median 27 learners/study)
  - mean MERSQI 13.3 (i.e. good)
  - technical (n = 44) and non-technical (n = 3) (e.g. communication, leadership) skills assessed
  - clinical areas included anesthesiology, endoscopy, resuscitation, and surgery
  - all studies included physicians/physicians-in-training
  - raters mainly (75%) physicians
- correlation between CL and GRS moderate
  - r = 0.76, 95% confidence interval [CI] 0.69–0.81
- inter-rater reliability was similar between scales
  - GRS 0.78, (95% CI 0.71–0.83), n = 23; checklist 0.81, (95% CI 0.75–
**Key Conclusions**

The authors conclude...

“[CL] inter-rater reliability and trainee discrimination were more favourable than suggested in earlier work, but each task requires a separate [CL]. Compared with the [CL], the GRS has higher average inter-item and inter-station reliability, can be used across multiple tasks, and may better capture nuanced elements of expertise.”

**Spare Keys – other take home points for clinician educators**

This is a great example of a planned research program leading to multiple wins. The original database developed for this systematic review has led to > 30 publications.