A systematic review of the reliability of objective structured clinical examination scores

Reference:
Brannick M¹, Erol-Korkmaz H², Prewett M¹. A systematic review of the reliability of objective structured clinical examination scores. Medical Education; 45(12):1181-9.

Author Institutions:
¹Department of Psychology, College of Arts and Sciences, University of South Florida, Tampa, Florida, USA
²Department of Psychology, Middle East Technical University, Ankara, Turkey

PubMed URL

Tags
Clinical domain
Medical Expert
Communicator

Educational domain
Assessment

Background

Objective Structured Clinical Examinations (OSCEs) have achieved widespread use, especially in undergraduate medical education. The operational design of this assessment instrument makes it attractive to educators who require an (relatively) inexpensive, time-efficient instrument that is appropriate for high-stakes judgments about learners.

The strict operational definition of an OSCE has evolved. A contemporary definition of an OSCE would include the following elements:
1. A circuit of focused clinical problems, not necessarily arranged in a related fashion
2. Learners simultaneously progress through stations
3. Some stations include standardized patients (i.e. actors) that simulate real patient presentations
4. Scoring includes direct observation by a clinical expert
5. Scoring tools use a variation of global response scores or checklists.

Purpose

This study purports to determine the reliability of an OSCE. Specifically, the authors pose three questions.
1. What reliability should be expected (on average) when developing an OSCE?
2. What is the likely range of such values?
3. What factors appear to influence the expected reliability?

**Type of paper**

Meta-analysis

**Key Points on the Methods**

The study provides limited detail regarding scope of the search, selection and quality scoring of studies. Key elements of the PRISMA standards for meta-analyses are missing. Specifically, a study selection flow diagram, a table of included study characteristics, a sensitivity analysis (for publication bias), and tests for statistical heterogeneity are not included.

While the authors acknowledge the inherent biases of Cronbach’s alpha for determining the reliability of an OSCE, it was adopted in this analysis because of the ubiquity with which it is reported in the literature.

Some studies reported between station reliability, while others reported within station reliability. The latter (optimistically) assumes that an assessor would objectively score related, but different items, pertaining to a brief clinical encounter.

No generalizability studies were identified that incorporated the variance both between and within stations.

**Key Outcomes**

The original search identified 98 studies and included 64. 39 studies reporting an alpha value were combined in the meta-analysis.

65% of studies involved medical students, 23% residents with the remaining a mix of these two groups or specific to physicians in practice.

Between stations, the unweighted alpha (overall reliability) equals 0.66 (95% CI 0.62-0.70). This study also confirms a well-accepted principle that reliability increases with the number of stations. The unweighted alpha for OSCEs with less than 10 stations was 0.56 and 0.74 for OSCEs with more than 10 stations. However, variability in this measure of reliability (even with a large number of stations) is still noted.

Reliability improved with 2 assessors versus 1 assessor and with expert assessors (i.e. clinician) versus standardized patient assessors.

**Key Conclusions**
The authors conclude...
“Overall scores on the OSCE are often not very reliable... It is generally helpful to use two examiners and large numbers of stations, but some OSCEs appear more reliable than others for reasons that are not yet fully understood.”

Recognizing that methodological limitations prevent making strong conclusions, contrary to the guarded conclusions of the authors, OSCEs are moderately reliable (as demonstrated here) and valuable in high-stakes examinations. Interestingly, the authors do not comment on the essential influence of context on psychometric influences. Medical educators should be cautioned that intangible local elements, reflected in their learners, assessors and stations, significantly influence the reliability of their particular OSCE.

One other note regards the endorsement of a checklist over a global response score (based on the differences in reliability) in this paper. This endorsement neglects the influence of halo/millstone effects of multiple measures and the de-valuing of progressive expertise that occur with checklists.

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

This paper is a good example of the challenges a “statistically-challenged” medical educator experiences in critiquing (and more importantly applying) the results of a paper. More significantly, it should be a reminder to seek out psychometric experts (i.e. reliability, validity) for guidance on assessment instrument design.