Published article:

Homer, MS, Fuller, R and Pell, G (2013) *Longitudinal interrelationships of OSCE station level analyses, quality improvement and overall reliability.* Medical Teacher, 35 (6). 515 - 517 . ISSN 0142-159X

http://dx.doi.org/10.3109/0142159X.2013.775415

Title Page

Title: Longitudinal interrelationships of OSCE station level analyses, quality improvement and overall reliability

Running Title: Station level metrics: Beneath the reliability iceberg

Authors:
Richard Fuller, Matthew Homer, Godfrey Pell
Leeds Institute of Medical Education
School of Medicine
University of Leeds
LS2 9JT
UK

Corresponding Author:
Richard Fuller
Leeds Institute of Medical Education
School of Medicine
University of Leeds
LS2 9JT
UK
Telephone: (+44) 0113 343106; Email: R.Fuller@leeds.ac.uk

Declaration of Interest
The authors have no declarations of interest

Notes on Contributors

Richard Fuller, MA, MBChB, FRCP is a consultant physician and Director of the Leeds MBChB undergraduate degree programme at Leeds Institute of Medical Education. His research interests focus on monitoring and improving the quality of assessment at programme levels, with particular focus on performance assessment

Godfrey Pell, BEng, MSc, FRSS, C.Stat, C.Sci is the senior statistician at Leeds Institute of Medical Education, who has a strong background in management. Before joining the University of Leeds, he was based in the Centre for Higher Education Practice at the Open University. His current research focuses on quality analysis within the OSCE and theoretical and practical applications of sequential test methods within performance testing.

Matthew Homer, BSc, MSc, PhD is a research fellow at the University of Leeds, working jointly across the Schools of Medicine and Education. His key healthcare related research focuses on the application of Item Response Theory in the analysis, quality management and development of written test items.

Summary

Objective Structured Clinical Examinations (OSCEs) are a key component within many healthcare assessment programmes. Quality assurance is designed to ensure rigour and credibility in decision making for both candidates and institutions, and most commonly expressed by a single measure of reliability. How overall reliability interrelates with OSCE station level analyses is less well established, especially in respect of the impact of quality improvements.

This work examined longitudinal interrelationships of reliability and station level metrics alongside interventions to improve the OSCE, revealing an anticipated relationship between poor reliability and poor station level analyses. However, longitudinal profiling revealed that overall reliability proved relatively unresponsive to continued improvements across stations – highlighting the importance of station level analyses as a routine part of any assessment quality assurance

Background

Across healthcare education, Objective Structured Clinical Examinations (OSCEs) are a key component of overall programmatic assessment structures. Originally conceptualised as highly structured assessment tools, initial OSCE and station design was often reductionist, but has undergone significant development, evolving as a valid, rigorous and sophisticated testing tool (Boursicot et al 2007). OSCEs are now successfully used to assess clinical performance in a wide range of settings and levels of clinical training.

The majority of OSCEs are used for high stakes testing, where it is essential that quality assurance allows fair, rigorous decision making about candidates and supports institutional credibility. Whilst the use of a single quality 'metric' across multiple forms of assessment provides limited inferences about quality, overall measures of reliability/internal consistency of an assessment are often the *only* measurement used to define quality (Brannick et al. 2011; Fuller et al. 2012). An increasing focus on how we define and use current thinking about validity with a programme of assessment challenges such 'one size' approaches to psychometric analyses (Schuwirth & van der Vleuten 2012).

Within OSCE settings, significant work has outlined the value of deeper, station level analyses of the outcomes of high stakes tests. This has the accompanying benefits of detecting previously unknown problems, designing appropriate solutions and then measuring impact of interventions (Pell at al 2010). There is an obvious benefit in always applying such analyses when overall reliability is demonstrably poor, but this approach does require additional expertise, time and cost - not always readily accessible in many healthcare educational settings. However, the evidence for undertaking such analyses to monitor OSCE improvements when overall reliability is apparently adequate is less clear.

This work set out to examine the longitudinal relationship of station level analyses with overall reliability, and the degree to which this reflected station level improvements.

Activity

We used longitudinal whole-exam and station-level data gathered from 2006-2009 from a final year qualifying OSCE (typically 18 stations with a total testing time of ~ 3 hours) in a UK undergraduate medical school. Standard setting was undertaken using the borderline regression method. [Pell & Roberts 2006]

The quality data routinely gathered comprised overall reliability (internal consistency) as determined by cronbach's alpha, and a range of Station Level Metrics (SLMs) including R^2 coefficient of determination (R is the correlation between checklist score and global rating), inter-grade discrimination, between group variance and 'alpha with item deleted' (Pell et al. 2010). This data is used to guide and evaluate further interventions to improve the OSCE, which have included improved assessor support (in addition to existing examiner training programmes) and revision of global grade descriptors. Checklist improvements have included work on anchors, item chunking to allow better construct alignment, and a focus on the fidelity of checklist items to ensure capture of higher level behaviours (e.g. safety) pertinent to station content (Pell et al 2010; Sadler 2010).

To examine the longitudinal interrelationship of overall reliability with these SLMs, and the impact of changes to OSCE design, we focused on three physical examination stations that were routinely blueprinted into the OSCE over the period 2006-2009 (Abdominal Examination, Cardiovascular Examination and Respiratory Examination). For each station, two important SLMs (R^2 and between group variance) were read against whole exam reliability and alongside major interventions applied to all final year OSCE stations

Results

The longitudinal relationship of SLMs with overall reliability is shown in Table 1. 2006 data indicates a poor reliability, typified at station level with low checklist score-global correlation (R^2) and high levels of between group variance (V%). A poor R^2 correlation would be below 0.5 and a >30-40% level of variance would indicate concern about the degree of construct irrelevant (i.e. error) variance.

	2006		2007		2008		2009	
	R^2	V%	R^2	V%	R^2	V%	R^2	V%
Abdominal Station	0.48	70	0.51	38	0.58	25	0.5	6
Cardio-Vascular Station	0.40	68	0.54	36	0.62	32	0.73	15
Respiratory Station	0.39	64	0.58	30	0.55	36	0.57	4
Overall OSCE reliability	α= 0.62		α= 0.75		α= 0.77		α= 0.76	

Table 1: Longitudinal relationship of station level metrics and overall reliability

The impact of considerable development of checklists and global grade descriptors was seen in the 2007 data, with an increase in cronbach's alpha and better SLMs for all stations. Further interventions followed after 2007 and 2008 OSCEs (chunking of lower level items and construct alignment to improve checklists).

Of particular note is that whilst SLMs continue to improve for these stations (and are representative of analyses for other stations used in OSCEs in these years), alpha appears unresponsive, suggesting that overall reliability is relatively insensitive to this continued station level improvement.

Discussion

Whilst overall reliability remains an important measurement of the quality of OSCE assessment formats, its limitations are well recognised (Fuller et al. 2012; Pell at al. 2010). The findings from this study support the association of a poor reliability with weaker SLMs, yet reveal that higher levels of reliability may not automatically follow from improvements of an individual station. At higher levels of reliability other factors such as more focused assessor training, may assume greater prominence. Importantly, institutional actions designed to improve the OSCE at the station level are not always apparent in global metrics. This degree of insensitivity may arise partly as a result of the natural between-station variation in student performance; the higher level items provide additional scope at the station level for good discrimination between candidates, but across the OSCE as a whole the borderline or weaker students still perform erratically, thereby limiting the overall reliability (White et al. 2008).

The multiple interventions to improve OSCE stations in this study were applied across our entire bank of stations rather than as a controlled ‘experiment’. Whilst we describe a number of interventions with resultant positive impact on quality (as measures by SLMs?), these are by no means generic or transferable in their entirety. Similarly, the work reports findings from one programme of study, and that quality improvement must be contextual (e.g. station content or overall OSCE design). Other factors will have contributed to the overall alpha, including other station level improvements and continued assessor training. However, the purpose of this work was not to examine the improvements per se, but the ability and limitations of psychometric analyses to capture the impact longitudinally.

Station Level Metrics are clearly an essential component of OSCE quality assurance, but these benefits have a wider impact (Pell et al. 2010). Measuring error variance allows an exploration of the degree to which this is construct relevant, the feasibility of station/checklist redesign or further assessor training, and a potential application across all criterion based testing formats within a programme of assessment,. Researching improvements at station level, and applying these systematically across the entire OSCE assessment (irrespective of context), is much more likely to improve quality, and provide a series of measures to quantify change.

References

Boursicot KAM, Roberts TE, Burdick WP. 2007. Structured Assessments of Clinical Competence. Understanding Medical Education Series, Association for the Study of Medical Education, Edinburgh

Brannick MT, Erol-Korkmaz HT, Prewett M. 2011. A systematic review of the reliability of objective structured clinical examination scores'. Medical Education. 45(12):1181-1189

Fuller R, Pell G, Homer M, Roberts T. 2012. Comments on 'A systematic review of the reliability of objective structured clinical examination scores'. Medical Education. 46(3):337

Schuwirth LTS, van der Vleuten CPM. 2012. Programmatic Assessment and Kane's validity perspective. Medical Education. 46(1): 38-48

Pell G, Fuller R, Homer M, Roberts TE. 2010. How to measure the quality of the OSCE: A review of metrics – AMEE guide no. 49. Med Teach 32:802-811

Pell G, Roberts TE. 2006. Setting standards for student assessment. Int J Res Method Educ. 29(1):91-103

Sadler DR. 2010. Fidelity as a preconception for integrity in grading academic achievement. Assessment & Evaluation in Higher Education. 35(6):727-743

White CB, Ross PT, Haftel MR. 2008. Assessing the Assessment: Are Senior Summative OSCEs Measuring Advanced Knowledge, Skills and Attitudes? Academic Medicine. 83(12):1191-1195