UNIVERSITY OF LEEDS

Article:
Homer, MS orcid.org/0000-0002-1161-5938, Pell, G, Fuller, R et al. (1 more author) (2016) Quantifying error in OSCE standard setting for varying cohort sizes: A resampling approach to measuring assessment quality. Medical Teacher, 38 (2). pp. 181-188. ISSN 0142-159X

https://doi.org/10.3109/0142159X.2015.1029898

Title page

Quantifying error in OSCE standard setting for varying cohort sizes: a resampling approach to measuring assessment quality

Running head: Quantifying error in OSCEs using resampling

Authors

Matt Homer*, Godfrey Pell*, Richard Fuller* and John Patterson+

Affiliations

* = Leeds Institute of Medical Education, University of Leeds

+ = Institute of Health Sciences Education, Barts and the London School,of Medicine and Dentistry, Queen Mary University of London

Corresponding author

Matt Homer

LIME, School of Medicine, University of Leeds, Worsley 7.09, Leeds LS2 9JT, UK

Telephone: +44 (0) 113 3434304. Fax: +44 (0) 113 3434375

Email: m.s.homer@leeds.ac.uk

Abstract

Background

The use of the borderline regression method (BRM) is a widely accepted standard setting method for OSCEs. However, it is unclear whether this method is appropriate for use with small cohorts (e.g. specialist post-graduate examinations).

Aims and methods

This work uses an innovative application of resampling methods applied to four pre-existing OSCE data sets (number of stations between 17 and 21) from two institutions to investigate how the robustness of the BRM changes as the cohort size varies. Using a variety of metrics, the 'quality' of an OSCE is evaluated for cohorts of approximately n=300 down to n=15. Estimates of the standard error in station-level and overall pass marks, R^2 coefficient, and Cronbach's alpha are all calculated as cohort size varies.

Results and conclusion

For larger cohorts (n>200), the standard error in the overall pass mark is small (less than 0.5%), and for individual stations is of the order of 1-2%. These errors grow as the sample size reduces, with cohorts of less than 50 candidates showing unacceptably large standard error. Alpha and R^2 also become unstable for small cohorts.

The resampling methodology is shown to be robust and has the potential to be more widely applied in standard setting and medical assessment quality assurance and research.

Key words

Borderline regression, bootstrapping, cohort size, error, OSCE, resampling, standard setting.

Introduction

The importance of quantifying error in standard setting

The measurement of 'quality' and detection and remediation of error is critical in any high stakes assessment as we seek to make defensible decisions about candidates based on a valid assessment (Kane, 2001). A range of psychometric indicators are typically used to 'assess assessment' and quantify error variance, and an increasing breadth of literature explores the contributory factors that underpin this, seeking to make sense of what was previously regarded as 'noise' and often not analysed (Govaerts, van der Vleuten, Schuwirth, & Muijtjens, 2007; van der Vleuten, 2014). As part of post hoc analysis, whole test reliability (Bland & Altman, 1997; Crossley, Davies, Humphris, & Jolly, 2003; Crossley et al., 2007; Tavakol & Dennick, 2011) and standard errors of measurement are calculated using classical test theory (McManus, 2012) or even item response approaches (Downing, 2003; Homer, Darling, & Pell, 2012; Lord, 1984), the latter utilised mainly in written assessments. As part of delivering a 'fair' OSCE, analysis can also correct for some problems post hoc, for example correcting for systematic site differences in multi-site exams (Pell, Fuller, Homer, & Roberts, 2010), or even adjusting for the systematic errors due to unwarranted assessor behaviour (Fuller, Homer, & Pell, 2011).

Despite this weight of literature, the estimation of the error in the standard setting itself has received relatively little attention, particularly in situations where generalizability theory (Bloch & Norman, 2012) may not be easily applicable; for example, in assessment where OSCE stations are not all scored on exactly the same scale.

If we accept that the outcomes of any examination are but one realisation from the universe of potential outcomes then this also holds for passing scores, regardless of how these are set - assuming the standard setting method is empirical or examinee-centred i.e. depends on a post-exam calculation using examinee outcomes from the assessment (Cizek & Bunch, 2007; Livingston & Zieky, 1982). On a different day, or with a different set of examiners, or a different set of students, or a different set of stations, the station-level and overall pass marks would very likely be different to those actually employed to make the pass/fail decisions. It would be very useful to be able to estimate how much variation (i.e. error, or more precisely, standard error) there is in these pass marks, and we might be re-assured if the error were shown to be 'small'. If however, this error were unacceptably large then this might indicate that a post hoc (examinee-centred) approach to standard setting were not appropriate in such a context.

How much does cohort size matter?

A key factor in the error in the setting of pass marks is the size of the cohort in question (i.e. the number of students sitting the exam), with greater levels of error typically associated with small cohorts. In such scenarios we have less post hoc

assessment data to make our standard setting decision on, so the error (i.e. standard error) in the pass mark will grow as cohort size decreases (Draper & Smith, 1998, p. 80; Muijtjens, Kramer, Kaufman, & Van der Vleuten, 2003; Wood, Humphrey-Murto, & Norman, 2006). This clearly has importance for institutions examining smaller cohorts, who need some guidance in accurately estimating the minimum number of candidates required for robust application of particular standard setting procedures in OSCEs.

The perspective from the literature

We review a number of studies in this section and find that whilst each provides useful additions to the literature, no single study contributes a comprehensive account of the substantive area of interest – the estimation of SEs in passing scores and quality metrics for OSCEs of varying cohort sizes.

In the borderline regression method (BRM) of standard setting (Kramer et al., 2003; Pell et al., 2010; Wood et al., 2006), checklist marks are regressed on global grades to estimate the weighted mean passing score in each station at the borderline grade. The overall OSCE passing score is then the average (mean) or sum of these station level passing scores, the latter approach is the one used in this study. Recognising the potentially adverse impact of assessment error, many institutions will also add a multiple of the standard error of measurement to prevent the occurrence of false positives (i.e. weaker students passing the OSCE through being *advantaged* by error) (Hays, Gupta, & Veitch, 2008)

In a study by Muijtjens *et al* (2003), a 'resampling' approach (using sub-samples of the original OSCE data without replacement) was used to estimate the error in OSCE BRM standard setting in two OSCEs of 16 and 11 stations with cohort sizes 86 and 155 candidates respectively. Through extrapolation methods, the focus of their work was on the estimation of the standard error at these full cohort sizes (found to be around 1.3% and 1.07% respectively). The analytic and standard setting methods employed in that paper focus on whole test level rather than including any station level analysis. The Muijtjens *et al* work (2003) also does not include estimates of other station level 'quality' metrics (e.g. Cronbach's alpha for internally consistency reliability and R^2 , the coefficient of determination, as a measure of the strength of the relationship between global grades and checklist scores within stations) and how these vary by cohort size (Fuller, Homer, & Pell, 2013; Pell et al., 2010).

A more recent study by Hejri et al (2013) quantified the error in the standard setting under the borderline regression method in a nine station OSCE sat by 105 candidates. They used a root-mean square error (RMSE) approach to estimate the error both at the station and overall OSCE level, but only for the fixed cohort size of 105. The overall error across the nine stations was estimated as 0.55%, with station-level errors of the order of 1-2% for a cohort of this size. The RMSE formula used for the calculations requires the total OSCE score to be computed as the average of the station scores, rather than, possibly, the total score across the OSCE. Perhaps more importantly, it also assumes that errors in pass marks are independent across stations. Analysis not included specifically in this paper suggests this latter assumption does not always hold in OSCE data – in other words for at least some

stations, pass mark errors are correlated. Ignoring such dependency will tend to systematically overestimate the error in the overall passing score.

Although a range of other work discusses and compares BRM to other standard setting methods, only a few papers highlight the estimation of error in standard setting itself in *small scale* OSCEs. For example, in a small cohort ten station OSCE (n=59), Wood *et al* (2006) calculate the standard error of the regression line within each station using an established regression-based formula (Draper & Smith, 1998, p. 80) and compare this to that of the modified borderline group (MBG) method (Cizek & Bunch, 2007, pp. 112–116; Wood et al., 2006). They find that within each station, BRM has smaller standard error (of the order of 2%, around 0.5% lower than MBG). However, this work does not give any error for standard setting in the overall OSCE, which is likely to be lower than at the station level - depending upon how the overall pass/fail decision is made. Also, it does not investigate how the error might change with cohort size and, as with the Muijtjens et al (2003) work cited earlier, variation or error in other metrics is not discussed. In another paper (Kramer et al., 2003), sub-sampling (as per the Muijtjens paper) is used to compare the Angoff and borderline regression methods of standard setting for a sixteen station OSCE of size n=86 (with resulting standard error 0.6%). The results are also extrapolated using generalizability theory to other cohort sizes but, again, errors at the station level and in other metrics are not discussed.

Other papers compare standard setting methods, sometimes in small scale OSCEs but do not calculate standard errors (Boursicot, Roberts, & Pell, 2007; Humphrey-Murto & MacFadyen, 2002; Schoonheim-Klein et al., 2009).

In summary, the literature as a whole indicates that BRM compares favourably with other methods of standard setting for OSCEs in terms of providing a robust and defensible standard, and there is some limited evidence that the variation in estimates of passing scores under BRM are smaller than those from other methods. However, we have been unable to find a comprehensive and systematic investigation into exactly how standard errors in pass marks (station and overall) and other related metrics vary by cohort size under the BRM.

This paper: more detailed quantifying of OSCE and station level error

In order to estimate the standard error of station-level and overall pass marks as set by the BRM, this paper analyses four sets of recent OSCE data from two medical schools and uses an innovative application of bootstrapping/resampling methods (Boos & Stefanski, 2010; Efron & Tibshirani, 1994; Wood, 2004). A key question in this research is how these errors in pass marks vary with cohort size – as cohort sizes get smaller, at what point do these errors become indefensibly large? What is the impact for institutions examining small cohorts in their choice of standard setting techniques?

Whilst the pass marks are obviously of key import in decision making, it is becoming established practice to evaluate the 'quality' of the assessment through the calculation of a range of post hoc metrics (often at the station level), rather than just relying on a single reliability metric (i.e. Cronbach's alpha, or a g-coefficient) (Fuller et al., 2013; Pell et al., 2010). As well as using all the global grade/checklist data, BRM affords computation of a wide range of these item and station level metrics –

for example, R^2 at the station level to indicate the extent to which the global grades and checklist marks are linearly related. In addition to pass mark standard errors, this work therefore also investigates errors in Cronbach's alpha and how these vary by cohort size, as well as looking at variation in R^2 at the station level.

In summary, this paper uses multiple sets of real OSCE data to investigate the robustness of the assessments under BRM standard setting, with a particular focus on quantifying standard errors in pass marks and other metrics (both at the exam and station-level) and how these vary with cohort size. There is a secondary, more methodological, purpose which is to illustrate the utility of resampling methods in the analysis of assessment quality, both within medical education and more broadly. We give some additional background to resampling methods in the next section.

Methods

A brief introduction to bootstrapping and resampling

Standard errors (SEs) are estimates, generated from a sample, of the uncertainty (i.e. spread) of a particular population parameter or statistic, often a mean (Altman & Bland, 2005). Often, distributional assumptions - for example normality of observations or of model errors (Draper & Smith, 1998, p. 34) - are required to facilitate the calculation of the SE using a closed formula but this is only possible in specific cases. Based on the usual assumptions of regression analysis (e.g. normality of errors) there is a formula for the SE (Draper & Smith, 1998, p. 80) at the station level. However, in the case of the overall pass mark, calculated in this study as the sum of station level pass marks (with station maxima not always the same),

there is no such closed formulae. Similarly for Cronbach's alpha, and many other commonly estimated statistics derived during a post-hoc analysis of OSCE data,. Note that we cannot easily apply a Generalizability approach (e.g. Raymond, Swygert, & Kahraman, 2012) - in the case where stations might be scored on different scales. Instead we turn to resampling methods as first introduced via bootstrapping by Efron in (1979) (see also Efron & Tibshirani, 1994). For basic introductions to bootstrapping and resampling see, for example, papers by Wood (2004) or Boos and Stefanski (2010).

The key premise of bootstrapping is that through sampling students *with replacement*, we actually have access to an unlimited number of 'samples' (usually of the same size as in the original cohort) hidden within the data. For each of these samples ('resamples') we can calculate any statistic that we like (e.g. pass marks, reliability figures) and hence can estimate the variation in each of these by looking at the resulting distribution of their values over a large number (often 1000) of resamples. The power of this approach is that repeated resampling from the sample mimics (hypothetical) resampling from the population, and hence gives otherwise unobtainable insight into the sampling variation of the statistic of interest (Boos & Stefanski, 2010; Wood, 2004). In addition, the properties required of the data under bootstrapping (e.g. distributional assumptions) are usually less restrictive compared to more conventional methods (Wood, 2004).

Data and sampling approaches

Table 1 shows the four sets of OSCE data that are used in the study.

TABLE 1 HERE

In each station a single assessor awards a checklist mark and an overall global grade.

For each OSCE data set, we resample the outcome data (i.e. student level marks and grades) 1,000 times (with replacement) for a fixed set of sample (i.e. 'cohort') sizes (n=15, 30, 50, 100, 150, 200, 250, up to the usual cohort size of between 250 and 300 – see Table 1). For each sample size and each iteration (i.e. each resample) we then calculate all the usual measures at both the station level (pass mark set using BRM and R^2, the amount of shared variance between checklist scores and global grades) and at the OSCE level (overall pass mark as sum of station pass marks, Cronbach's alpha). This gives us a set of 1,000 sets 'OSCE' data and we can see how much each of the metrics and pass marks vary over this thousand. By comparing between 'cohort' sizes we obtain clear insight as to how all these metrics and pass marks vary by cohort size.

We use R version 3.1.0 (2013) to carry out the main analysis – for an OSCE with 20 stations there are 20 (stations) × 1000 (resamples) × 8 (sample cohort sizes) =160,000 regression calculations to carry out in one resampling analysis. In order to check the stability of the results, each set of 1000 resamples for each OSCE has been run at least twice to produce multiple 'realisations' of the results. In each case, we find very similar results to those presented in this paper indicating that resampling a thousand times produces sufficiently robust (i.e. reproducible results) for this type of data – in the results section to follow we selectively demonstrate this by comparing two realisations.

Results

We focus in detail on one of our four OSCEs, Exam A, taken by third year students within a 5 year undergraduate medical programme (Table 1), investigating the error (i.e. the SE) in the overall pass mark, station level pass mark, Cronbach's alpha and R^2 across the 1000 resamples. This assessment consists of a range of stations, each with a particular focus (clinical and examination skills, and history taking). All of this analysis is carried out for a range of 'cohort' sizes to simulate successively smaller cohorts. We then give a summary comparison of results across all the four OSCE data sets analysed.

Standard errors of parameter estimates – Exam A

We look at how much the mean estimates for the pass marks and other metrics *vary* across the 1000 resamples as the cohort size decreases. Figure 1 shows that the SE of the total pass mark varies approximately by the inverse square root of the sample size (two realisations shown, both with R^2 figures over 0.99 indicating extremely good fit). One might expect this asymptotic behaviour based on standard statistical sampling theory (Rowntree, 2000, p. 94).

At the full cohort size, the SE in percentage terms is 0.36% – this is good evidence that the standard setting is quite robust at this value of n (i.e. a 95% confidence interval for the pass mark has half-width of approximately twice this=0.72%). However, at the lower cohort sizes, the standard errors do grow – to around 1.7% at n=15.

FIGURE 1 HERE

The qualitative nature of the graphs for station-level pass marks are similar (i.e. negative exponentials – see Figure 2) but, importantly, have larger percentage standard errors (from 1-2% at the largest sample size to 4-10% at n=15). To aid clarity, Figure 2 only shows a sub-set of four stations from OSCE A, including the stations with the smallest and largest SE for this OSCE (stations 11 and 5 respectively).

FIGURE 2 HERE

Alpha and R^2 show a similar pattern (Figures 3 and 4) but the SEs are much larger (note the vertical scales in this figures in comparison to those of Figures 1 and 2):

FIGURE 3 HERE

The percentage on the vertical scale in Figure 3 is based on maximum alpha=1.

FIGURE 4 HERE

The percentage on the vertical scale in Figure 4 is based on maximum R^2=1.

Comparisons across OSCE data sets

Table 2 shows a summary of the resampling results for each of the four data sets outlined in Table 1.

TABLE 2 HERE

The four analyses show a strong level of consistency in the overall results. The SE in the overall pass mark is of the order of 0.3% for the full cohort but rises to around 1.5% at n=15. Station level pass marks have an SE approximately four times as large as these whilst the corresponding estimates for Cronbach's alpha and R^2 are approximately nine and 14 times as large respectively.

It is important to note that we have used 'raw' data in all our analysis, with each station-level checklist score distribution having a different mean (and standard deviation). As part of a related study, we carried out a parallel analysis using standardised scores (i.e. all station score distributions set to have the same mean and standard deviation). We did this in order to investigate the impact of such a standardisation on the estimates of the SEs (as per Table 2), perhaps expecting that the SEs for such data would be smaller. In fact, whilst measures of reliability (alpha) are slightly higher for such standardised data, we have found that it makes no systematic difference to the SE results presented here.

Discussion

The overall OSCE

The primary aim of this study is to investigate how the error in pass marks and quality metrics vary with cohort size under the BRM of standard setting for OSCEs. For larger cohorts (n~250-300), the analysis shows that the overall pass mark as set using BRM method does not vary much between resamples – with a SE of the order of 0.3% for a 'standard' OSCE with 20 or so stations – with less stations this figure

would increase and vice versa. This is good evidence that the BRM is robust at these sample sizes with this number of stations, and these findings complement other research comparing the BRM favourably with other standards setting methods (Kramer et al., 2003; Schoonheim-Klein et al., 2009; Wood et al., 2006). The substantive findings are broadly similar across the two institutions involved, which gives some confidence as to their wider applicability.

In some institutions it is common practice to calculate the standard error of measurement (SEM) for the OSCE (Hays et al., 2008; Pell et al., 2010) which, under classical test theory, depends on the reliability (alpha or similar) (Streiner & Norman, 2008, pp. 191–192) – high reliability corresponds to low SEM, and vice versa. The SEM is a measure of uncertainty in the observed student score, and it (or a multiple thereof) is then added to the pass mark in order to minimise the possibility of false positive decisions (i.e. poor students passing the exam based on the error in the assessment scoring process acting in their favour). Typically, the SEM is of the order of 2-3% in our OSCEs (based on a standard deviation of scores of 6% and alpha=0.8). The point here is that the work presented in this paper demonstrates that the SE in the standard setting for the overall OSCE (for the full cohort) is an order of magnitude smaller than the error in the observed student score (i.e. the SEM). Hence, we can be confident that the SE in the standard setting is of relatively little consequence in terms of its impact on overall OSCE pass/fail decisions at these ‘large’ cohort sizes.

The initial motivation for this study was to investigate precisely how small cohort sizes can be whilst still providing acceptable, defensible pass marks under the BRM.

Our work quantifying the SE in the pass mark indicates that with cohorts of 50 candidates or less these start to become unacceptably large – of the order of 1% - in other words, approaching the same order magnitude as of the SEM itself (which isn't itself dependent on cohort size). In addition, our work suggests that other metrics such as alpha and (to a lesser extent) R^2 also become unstable for small cohorts (full details not included). One potential way of overcoming these difficulties at smaller institutions would be to increase the number of stations in the exam. However, this is known to be relatively inefficient since doubling the length of a test only increases reliability from (say) 0.70 to 0.82 (using the Spearman-Brown formula, Streiner & Norman, 2008, p. 88). Hence, for reasons of cost, logistics and redundancy, lengthening the 'test' sufficiently to bring error down sufficiently is likely to be impractical in most contexts. Where an OSCE station bank exists and the same station has been used more than once in an unaltered form, it may be possible to aggregate the data across test administrations and hence form a larger dataset for standard setting calculations. A final possibility is to move away from an examinee-centred standard setting approach to one that is test- or item-centred (e.g. Angoff or Ebel), where any error in pass marks does not depend on student cohort size. However, this move would bring its own difficulties as there is good evidence that in the assessment of complex behaviours as in the OSCE, the observation and scoring of actual performance, including the awarding of holistic judgements (i.e. global grades), produces more valid outcomes than does any *a priori* judgments of checklist item/station difficulty is likely to do (Kane, 2001).

Station level effects

As one might expect, station level pass marks have greater SE in percentage terms than the overall pass mark (of the order of 1-2% at the full cohort size – similar to the

estimates found in Hejri et al (2013)) - essentially 'outlier' assessor judgments have a proportionately greater effect on the pass mark in a single station than they do for the full OSCE. For medical schools where pass/fail decisions rest solely on the overall OSCE pass mark these station-level decisions are of relatively little import. But for schools that do have additional passing criteria (e.g. a requirement to pass a minimum number of stations or particular types of stations (Clauser, Clyman, Margolis, & Ross, 1996) this problem is of more concern. In this work, we have quantified the uncertainty in these decisions, and one might argue that this provides evidence in favour of a parsimonious approach to OSCE pass/fail decisions – using only the overall pass mark across the whole OSCE. If we are confident that we are successfully measuring 'clinical skills' *only* at the whole-OSCE level, a narrow psychometric viewpoint might argue that pass/fail decisions must be made based on this. Making decisions that rest, to an extent, on subsets of stations where decisions are subject to greater, possibly unacceptable, error could be construed as less robust (Sinharay, Haberman, & Puhan, 2007; Sinharay, Puhan, & Haberman, 2011). Another line of argument in favour of a single pass/fail OSCE standard is that in a more holistic or programmatic assessment framework the issue of compensation within a single 'test' becomes less important provided there are other instruments or tests that complement the OSCE in the appropriate areas (Dijkstra, Vleuten, & Schuwirth, 2010; Schuwirth & Vleuten, 2006).

However, these arguments must be tempered by a judgement about the maturity of the assessment programme within institutions. Further, the consequence of the generation of more 'discipline focused' test material are likely to be more reductionist, and counter to more integrated assessment. Other counter-arguments

to the single hurdle approach to OSCEs are more pragmatic: depending on the year group being assessed, stations are often classified into groups measuring particular types of behaviours (e.g. ‘skills’, ‘communication’, or perhaps individual specialties). An ongoing concern is that any assessment progression decisions should have fidelity with more ‘holistic’ ability. To avoid excessive compensation across these different domains additional sub-OSCE ‘hurdles’ must be a necessity through a conjunctive assessment model (Reece, Chung, Gardiner, & Williams, 2008; Sadler, 2009). If not, then some students might well pass the OSCE but, for example, fail the majority of stations, or fail stations in a particular domain.

Neither of these contrasting views is entirely congruent with many of the challenges within modern assessment – and, in particular, the impact of learner appeals on assessment behaviour by institutions. In many contexts, there remains a strong necessity in being able to ‘prove’ that decisions made in individual assessments are defensible.

Methodological issues, study limitations and future work

Over the course of this work, the resampling methodology employed has proved robust, producing replicable results with 1000 resamples across all cohort sizes. Whilst the technical challenge of carrying out the repeated calculations through a programing approach in R is initially considerable, the resampling approach is conceptually simple once one is prepared to accept that a single set of assessment data actually contains an very large number of ‘samples’ (formally, resamples) for a fixed cohort size (Wood, 2004). The method also rests on less restrictive assumptions than do alternative approaches based on arguably more challenging

RMSE and/or generalizability formulae. The substantive findings are broadly similar across institutions, which gives some confidence as to their wider applicability.

The key limitation of the study is that the findings with regard to smaller cohort sizes are not actually based on data from small institutions. The question as to how the results based on such data would compare to those presented here is obviously acute. Certainly a smaller institution is less likely to suffer from the error due to the large number assessors/circuits necessary in a larger school OSCE. Additional work using single circuits from our data indicates that our results might be over-stating the SE by about 20% for schools with small cohorts with a only single circuit (and, possibly, multiple sessions). However, whilst the SE in standard setting might be less in a single circuit institution, there is additional risk that assessor judgements at a station are not subject to the normative triangulation that is possible when comparing across circuits in a larger medical school.

The relative homogeneity of the candidate pool is another factor that might impact on the error, as some small cohorts taking specialist assessments are likely to be more homogeneous in this regard compared to larger (e.g. undergraduate) cohorts. However, despite these limitations, we feel the broad findings of our study are likely to generalise to other contexts.

Future work may include more complex two-step selection approaches (i.e. repeatedly resampling from a pre-selected sample of a fixed size) since this might more realistically ‘mimic’ small cohort data. It is also hoped that smaller institutions might make available their data for analysis to compare with our work – we would be

happy to share our R code to facilitate this. There is also the possibility of further investigation into circuit effects through resampling at the circuit level, and of resampling by station instead of by student. The sensitivity of all these analyses to the number of stations in the assessment and the degree of skew in the score distributions might also be investigated. Finally, there may be applications of resampling methods to other standard setting approaches to see how the standard errors and other metrics might compare to those presented here.

Conclusion

Advances in standard setting and an increasing sophistication of post hoc analyses allow institutions unparalleled ability to 'assess assessment'. Such analyses, coupled with emergent work that seeks to understand the impact of individual examiner behaviour (Gingerich, Regehr, & Eva, 2011; Kogan, Conforti, Bernabeo, Iobst, & Holmboe, 2011), OSCE design issues and context specificity also allow us to better define error and variance within our high stakes assessments. However, this improved understanding also brings challenges as institutions grapple with the defensibility and rigour of decision making alongside unanswered questions about the nature and impact of error estimation, particularly in small cohort OSCEs.

What then are institutions, particularly those with smaller cohorts, to conclude from this work? Our key message is that if possible, the errors in the standard setting should be estimated, conceivably using methods exemplified in this paper, and steps taken to ensure defensible pass/fail decisions are made, perhaps through increasing the pass-mark by an additional amount related to the total error identified (i.e. error in pass mark plus error in observed score). Modelling of the impact of such an

approach can be investigated using pre-existing data, and we have identified that cohorts of less than 50 candidates are associated with unacceptable levels of error in pass marks. It is clear that institutions need to carefully consider whether their standard setting methods are sufficiently robust, based both on empirical work and the latest research literature.

Practice points

- Using resampling methods one can calculate the error (standard error) in the pass mark calculations under examinee-centred methods of standard setting such as the borderline regression method (BRM).
- For large cohorts (e.g. 200+), the BRM pass mark has only a small standard error (<0.5%), providing additional validation of the BRM as a defensible methods of standard setting.
- The standard error in the pass mark grows as the cohort size decreases, and, particularly at the station level, becomes unacceptably large for small cohorts (e.g. n<50)
- Small intuitions should take care to estimate standard errors in pass marks and ensure that their pass/fail decisions under the BRM are sufficiently defensible.
- Resampling methods provide robust and conceptually straightforward ways to calculate (standard) error in a range of post hoc metrics.

Notes on contributors

MATT HOMER, BSc, MSc, PhD, CStat is a senior researcher and teacher at the University of Leeds, working in both the Schools of Medicine and Education. His

research generally has a quantitative methodological focus, and within medical education relates to evaluating and improving assessment quality, standard setting and psychometrics.

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. His current research focuses on quality within the OSCE, including theoretical and practical applications. He acts as an assessment consultant to a number of medical schools.

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.

JOHN PATTERSON BSc, PhD was formerly Associate Dean for Undergraduate Medical Studies and Head of MBBS Assessment at Bart's and the London School of Medicine and Dentistry. Since retirement, he has become an independent medical assessment consultant, providing workshops, assessment advice and analyses for a various HEIs, postgraduate colleges and regulatory bodies.

Acknowledgement

Thanks to Rob Long from the University of Leeds for the help in writing the R code used in this study.

Declarations of interest

The authors report no declarations of interest.

Glossary

Resampling – this is a process of drawing repeated samples with replacement from a sample to estimate the precision (i.e. variation) of a particular statistic (e.g. a mean – for each drawn sample, the mean is calculated and then the spread in distribution of these means gives a measure of variation of the population mean).

References

Altman, D. G., & Bland, J. M. (2005). Standard deviations and standard errors. *BMJ*, *331*(7521), 903. doi:10.1136/bmj.331.7521.903

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *BMJ*, *314*(7080), 572. doi:10.1136/bmj.314.7080.572

Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Medical Teacher*, *34*(11), 960–992. doi:10.3109/0142159X.2012.703791

Boos, D., & Stefanski, L. (2010). Efron's bootstrap. *Significance*, *7*(4), 186–188. doi:10.1111/j.1740-9713.2010.00463.x

Boursicot, K. A. M., Roberts, T. E., & Pell, G. (2007). Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Medical Education*, *41*(11), 1024–1031. doi:10.1111/j.1365-2923.2007.02857.x

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting a guide to establishing and evaluating performance standards on tests*. Thousand Oaks, Calif.: Sage Publications. Retrieved from http://SRMO.sagepub.com/view/standard-setting/SAGE.xml

Clauser, B. E., Clyman, S. G., Margolis, M. J., & Ross, L. P. (1996). Are fully compensatory models appropriate for setting standards on performance assessments of clinical skills? *Academic Medicine: Journal of the Association of American Medical Colleges*, *71*(1 Suppl), S90–92.

Crossley, J., Davies, H., Humphris, G., & Jolly, B. (2003). Generalisability: a key to unlock professional assessment. *Medical Education*, *37*(6), 574–574. doi:10.1046/j.1365-2923.2003.01546.x

Crossley, J., Russell, J., Jolly, B., Ricketts, C., Roberts, C., Schuwirth, L., & Norcini, J. (2007). “I”m pickin’ up good regressions’: the governance of generalisability analyses. *Medical Education*, *41*(10), 926–934. doi:10.1111/j.1365-2923.2007.02843.x

Dijkstra, J., Vleuten, C. P. M. V. der, & Schuwirth, L. W. T. (2010). A new framework for designing programmes of assessment. *Advances in Health Sciences Education*, *15*(3), 379–393. doi:10.1007/s10459-009-9205-z

Downing, S. M. (2003). Item response theory: applications of modern test theory in medical education. *Medical Education*, *37*(8), 739–745.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis*. New York: Wiley. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=26118

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. doi:10.1214/aos/1176344552

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Fuller, R., Homer, M., & Pell, G. (2011). *What a difference an examiner makes! __Detection, impact and resolution of “rogue” examiner behaviour in high stakes OSCE assessments*. Presented at the AMEE, Vienna.

Fuller, R., Homer, M., & Pell, G. (2013). Longitudinal interrelationships of OSCE station level analyses, quality improvement and overall reliability. *Medical Teacher*, *35*(6), 515–517. doi:10.3109/0142159X.2013.775415

Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-Based Assessments as Social Judgments: Rethinking the Etiology of Rater Errors: *Academic Medicine*, *86*, S1–S7. doi:10.1097/ACM.0b013e31822a6cf8

Govaerts, M. J. B., van der Vleuten, C. P. M., Schuwirth, L. W. T., & Muijtjens, A. M. M. (2007). Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Advances in Health Sciences Education: Theory and Practice*, *12*(2), 239–260. doi:10.1007/s10459-006-9043-1

Hays, R., Gupta, T. S., & Veitch, J. (2008). The practical value of the standard error of measurement in borderline pass/fail decisions. *Medical Education*, *42*(8), 810–815. doi:10.1111/j.1365-2923.2008.03103.x

Hejri, S. M., Jalili, M., Muijtjens, A. M. M., & Van Der Vleuten, C. P. M. (2013). Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*, *18*(10), 887–891.

Homer, M., Darling, J., & Pell, G. (2012). Psychometric characteristics of integrated multi-specialty examinations: Ebel ratings and unidimensionality. *Assessment & Evaluation in Higher Education*, *37*(7), 787–804. doi:10.1080/02602938.2011.573843

Humphrey-Murto, S., & MacFadyen, J. C. (2002). Standard setting: A comparison of case-author and modified borderline-group methods in a small-scale OSCE. *Academic Medicine*, *77*(7), 729–732. doi:10.1097/00001888-200207000-00019

Kane, M. (2001). So Much Remains the Same: Conception and Status of Validation in Setting Standards. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives*. Lawrence Erlbaum Associates.

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: a conceptual model. *Medical Education*, *45*(10), 1048–1060. doi:10.1111/j.1365-2923.2011.04025.x

Kramer, A., Muijtjens, A., Jansen, K., Düsman, H., Tan, L., & Van Der Vleuten, C. (2003). Comparison of a rational and an empirical standard setting procedure for an OSCE. *Medical Education*, *37*(2), 132–139. doi:10.1046/j.1365-2923.2003.01429.x

Livingston, S. A., & Zieky, M. J. (1982). Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Educational Testing Service, Box 2885, Princeton, NJ 08541 ($7.50). Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED227113

Lord, F. M. (1984). Standard Errors of Measurement at Different Ability Levels. *Journal of Educational Measurement*, *21*(3), 239–243. doi:10.1111/j.1745-3984.1984.tb01031.x

McManus, I. C. (2012). The misinterpretation of the standard error of measurement in medical education: a primer on the problems, pitfalls and peculiarities of the three different standard errors of measurement. *Medical Teacher*, *34*(7), 569–576. doi:10.3109/0142159X.2012.670318

Muijtjens, A. M. M., Kramer, A. W. M., Kaufman, D. M., & Van der Vleuten, C. P. M. (2003). Using Resampling to Estimate the Precision of an Empirical Standard-Setting Method. *Applied Measurement in Education*, *16*(3), 245–256. doi:10.1207/S15324818AME1603_5

Pell, G., Fuller, R., Homer, M., & Roberts, T. (2010). How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. *Medical Teacher*, *32*(10), 802–811. doi:10.3109/0142159X.2010.507716

Raymond, M. R., Swygert, K. A., & Kahraman, N. (2012). Measurement precision for repeat examinees on a standardized patient examination. *Advances in Health Sciences Education: Theory and Practice*, *17*(3), 325–337. doi:10.1007/s10459-011-9309-0

R Core Team. (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/.

Reece, A., Chung, E. M. K., Gardiner, R. M., & Williams, S. E. (2008). Competency domains in an undergraduate Objective Structured Clinical Examination: their impact on compensatory standard setting. *Medical Education*, *42*(6), 600–606. doi:10.1111/j.1365-2923.2008.03021.x

Rowntree, D. (2000). *Statistics Without Tears: An Introduction for Non-Mathematicians*. London: Penguin Books.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, *34*(2), 159–179. doi:10.1080/02602930801956059

Schoonheim-Klein, M., Muijtjens, A., Habets, L., Manogue, M., van der Vleuten, C., & van der Velden, U. (2009). Who will pass the dental OSCE? Comparison of

the Angoff and the borderline regression standard setting methods. *European Journal of Dental Education*, *13*(3), 162–171. doi:10.1111/j.1600-0579.2008.00568.x

Schuwirth, L. W. T., & Vleuten, C. P. M. (2006). A plea for new psychometric models in educational assessment. *Medical Education*, *40*(4), 296–300. doi:10.1111/j.1365-2929.2006.02405.x

Sinharay, S., Haberman, S., & Puhan, G. (2007). Subscores Based on Classical Test Theory: To Report or Not to Report. *Educational Measurement: Issues and Practice*, *26*(4), 21–28. doi:10.1111/j.1745-3992.2007.00105.x

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME Instructional Module on Subscores. *Educational Measurement: Issues and Practice*, *30*(3), 29–40. doi:10.1111/j.1745-3992.2011.00208.x

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: a practical guide to their development and use* (4th ed.). Oxford; New York: Oxford University Press.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. doi:10.5116/ijme.4dfb.8dfd

Van der Vleuten, C. P. M. (2014). When I say … context specificity. *Medical Education*, *48*(3), 234–235. doi:10.1111/medu.12263

Wood, M. (2004). Statistical inference using bootstrap confidence intervals. *Significance*, *1*(4), 180–182. doi:10.1111/j.1740-9713.2004.00067.x

Wood, T. J., Humphrey-Murto, S. M., & Norman, G. R. (2006). Standard Setting in a Small Scale OSCE: A Comparison of the Modified Borderline-Group Method and the Borderline Regression Method. *Advances in Health Sciences Education*, *11*(2), 115–122. doi:10.1007/s10459-005-7853-1

Figure legends

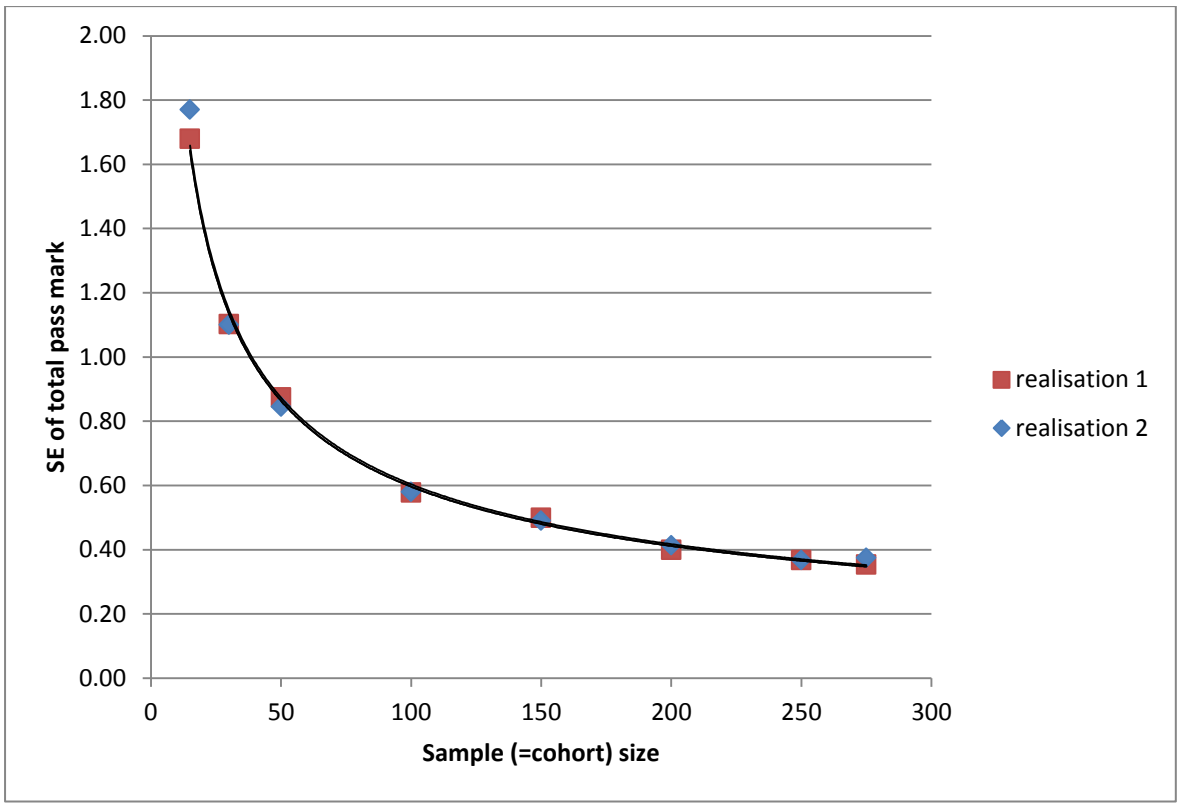
Figure 1: Standard error of overall OSCE pass mark by cohort size (two realisations).

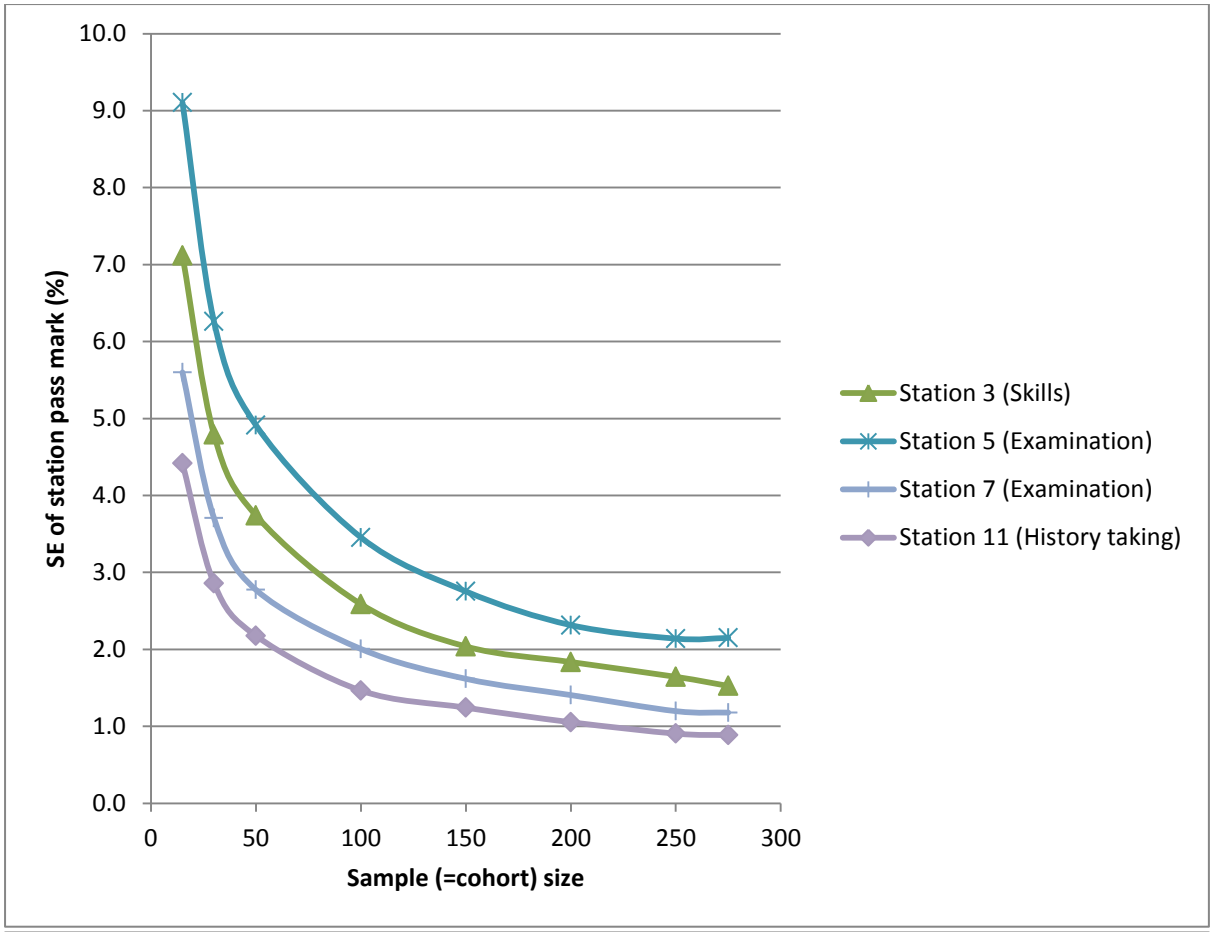
Figure 2: Standard errors for select station level pass marks by cohort size

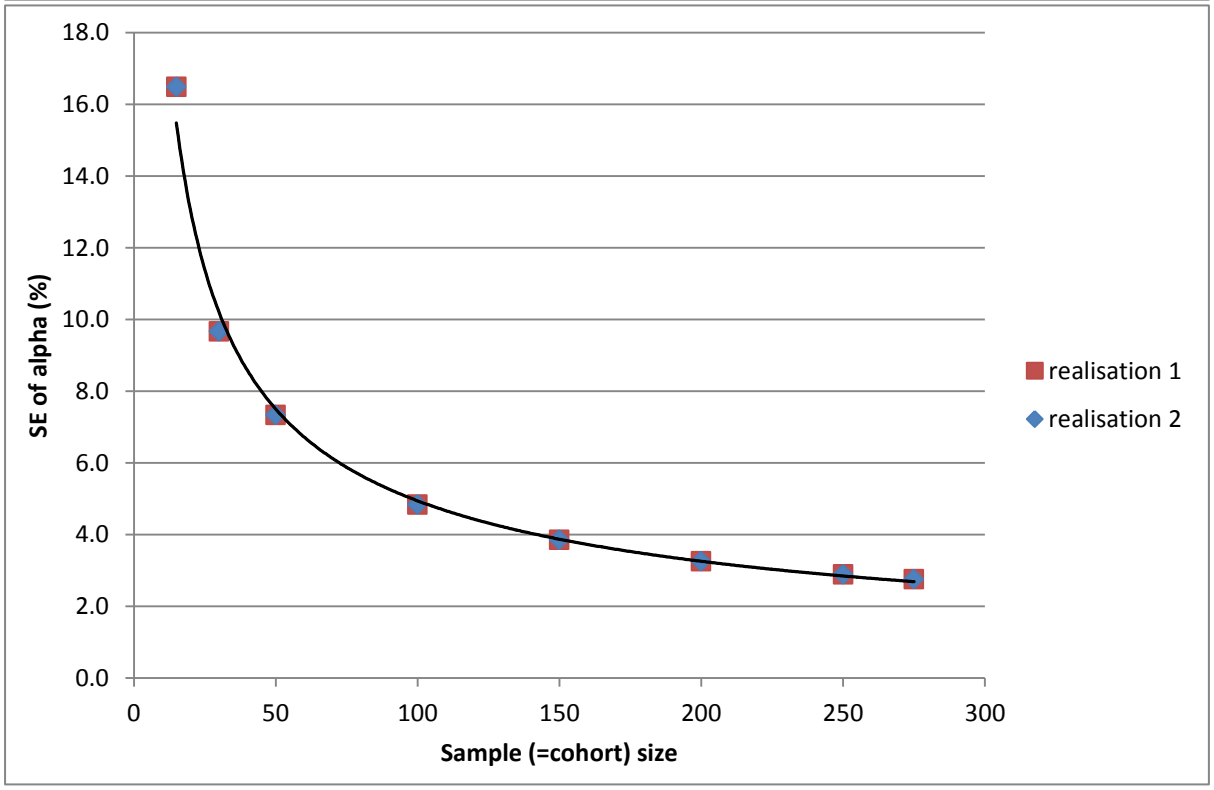
Figure 3: Standard error of Cronbach's alpha by cohort size (two realisations)

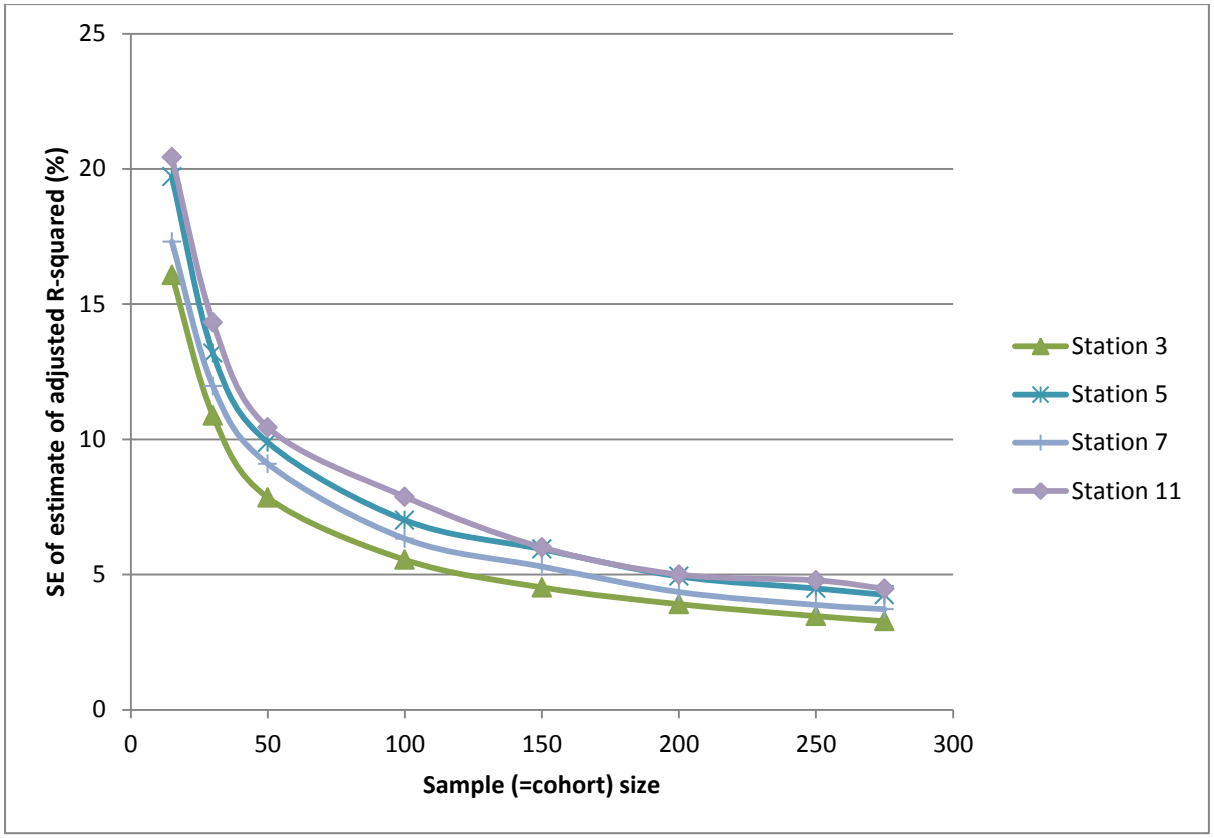
Figure 4: Standard error of R^2 *for select stations by cohort size*

Figures


2.00
1.80
1.60
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
SE of total pass mark
0
50
100
150
200
250
300
Sample (=cohort) size
realisation 1
realisation 2


10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
SE of station pass mark (%)
0
50
100
150
200
250
300
Sample (=cohort) size
Station 3 (Skills)
Station 5 (Examination)
Station 7 (Examination)
Station 11 (History taking)


18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
SE of alpha (%)
0
50
100
150
200
250
300
Sample (=cohort) size
realisation 1
realisation 2


25
20
15
10
5
0
SE of estimate of adjusted R-squared (%)
0
50
100
150
200
250
300
Sample (=cohort) size
Station 3
Station 5
Station 7
Station 11

Tables

Exam	Year of exam	Year group	Number of students	Number of stations	Checklist marks per station	Global grades in each station	Total marks in OSCE
A	2012	3	275	21	19 to 37	0=fail 1=borderline 2=clear pass 3=good pass 4=excellent pass	642
B	2012	4	282	20	23 to 39		611
C	2010	5	272	18	31 to 47		680
D	2012	5	327	17	10	0 = Fail 1= Borderline 2 = Pass 3 = Good pass	100

Table 1: OSCE data used in the study

Exam	Year of exam	Year group	Standard error (%)											
			Overall pass mark			Mean Station level pass mark			Cronbach's alpha			Mean R^2		
			Full cohort	n=50	n=15	Full cohort	n=50	n=15	Full cohort	n=50	n=15	Full cohort	n=50	n=15
A	2012	3	0.36	0.85	1.70	1.47	3.49	6.82	2.76	7.34	16.5	4.14	9.79	19.0
B	2012	4	0.30	0.72	1.36	1.08	2.58	4.93	2.54	6.50	15.5	4.43	10.42	19.7
C	2010	5	0.38	0.89	1.69	1.21	2.84	5.57	2.43	5.92	15.3	4.05	9.65	18.9
D	2012	5	0.26	0.67	1.29	1.09	2.84	5.22	2.25	6.31	14.5	4.62	9.97	19.3

Table 2: Standard errors as a percentage across the four OSCE data sets