Published article:
Homer, MS, pell, G and fuller, R (2012) *Is short term remediation after OSCE failure sustained? A retrospective analysis of the longitudinal attainment of underperforming students in OSCE assessments.* Medical Teacher, 34 (2). 146 - 150 . ISSN 0142-159X
http://dx.doi.org/10.3109/0142159X.2012.643262

Title

Is short term remediation after OSCE failure sustained?
A retrospective analysis of the longitudinal attainment of underperforming students in OSCE assessments

Short title

Is short term OSCE remediation sustained?

Names of authors

G. Pell, R. Fuller, M. Homer

Institution

Leeds Institute of Medical Education, School of Medicine, University of Leeds, LS2 9NL

Corresponding author

G. Pell, Leeds Institute of Medical Education, School of Medicine, University of Leeds, LS2 9NL, 0113 3434378, g.pell@leeds.ac.uk

Abstract

Background

Significant improvements in the delivery of criterion based assessment techniques have improved confidence in standard setting and assessment quality. However, for underperforming students, a lack of evidence about longitudinal performance of this group poses dilemmas to educators when making decisions about the timing and nature of remediation.

Aims

To investigate the longitudinal performance of the UK undergraduate medical degree students, with a particular focus on comparing the poorly performing students (i.e. those with borderline or failing grades) with the main cohort of students.

Method

Over a five year period, 3200 student OSCE assessments from a single medical school consisting were investigated. A poorly performing subgroup of 125 students was identified and their longitudinal performance in the final three years of the undergraduate medical degree was analysed.

Results

The relative performance of this student group declines across serial OSCEs, despite current methods of 'remediation and re-test'.

Conclusions

This analysis demonstrates that typically students in the poorly performing subgroup achieve only short term success with traditional remediation and re-test models, and critically show an absence of longitudinal improvement. There is a clear need for institutions to develop profiling models that can help identify this student group and develop effective, research led models of remediation.

Practice points

- Under the usual system of remediation (assessment focussed revision program and then re-assessment) the majority of poorly performing students fail to improve in clinical assessments

- Following poor performance, remediation should be embedded in the subsequent program

Notes on contributors

Godfrey Pell is a senior statistician who has a strong background in management. Current research includes standard setting and quality for practical assessment in higher education, especially OSCEs.

Dr Richard Fuller is a Consultant Geriatrician, and Director of the MBChB at Leeds. His research interests include clinical assessment, in particular monitoring and improving the quality of the OSCE.

Dr Matthew Homer is a Research Fellow at the University of Leeds, working in the both the Schools of Medicine and Education. He works on a range of research projects and provides general statistical support to colleagues. His research interests include the statistical side of assessment, particularly related to OSCEs.

Professor Trudie Roberts is Director of both the Leeds Institute of Medical Education and the Medical Education Unit in the School of Medicine at the University of Leeds.

Her main interests and expertise are in the areas of assessment of competence, professionalism, inter-professional education and widening access and participation.

Introduction

Within undergraduate medical education, considerable investment is made to ensure students reach and maintain desired standards, and to ensure assessment processes demonstrate sufficient rigour in setting appropriate performance standards (Cizek & Bunch 2007; Streiner & Norman 2008). Consequently, methods of assessment and standard setting must be defensible when subject to detailed scrutiny (Regehr et al. 1998; Roberts et al. 2006; Cohen et al. 2009). Delivery of a high stakes clinical assessments are complex and costly and many standard setting methodologies are labour intensive (such as Angoff and Ebel), requiring panels of experts to consider all questions in advance, or necessitate the analysis of large data sets using techniques such as borderline, or contrasting group methods (Cusimano 1994; Cusimano et al. 1995; Cizek & Bunch 2007).

Similarly, there is a clear responsibility for institutions to ensure underperforming medical students are detected, remediated and supported. There is an analogous requirement to undertake rigorous re-assessment to permit progression throughout courses and satisfy regulatory requirements for fitness to practice. Although considerable time and effort is devoted to the topic of student assessment, little research specifically addresses the issue of students who underperform and resit (Ricketts 2010).

Studies using early OSCE data highlighted that under-performance in OSCEs undertaken in the third year of an undergraduate medical programme was strongly associated with later poor performance, although reported reliability in these examinations was low (cronbach's α 0.68), with no station level analysis (Martin & Jolly 2002). Other work focusing on remediation after underperformance revealed expected, and wide ranging academic and non-academic difficulties amongst failing students that provided opportunities for remediation (Sayer et al. 2002).

A recent paper has addressed the issue of the relative difficulty of resit assessments when compared to the main assessment, finding that students who performed poorly in the main assessment seemed to have little trouble in the resit (Pell et al. 2009). Other studies have increasingly highlighted short-term successful remediation after OSCE failure, or the use of the assessment for learning models to attempt to detect, and remediate poor performance in advance of high-stakes assessment (White et al. 2009; Cleland et al. 2010)

However, these studies have focused on a model of test-remediate-retest, with lack of longitudinal data. There have been legitimate criticisms that the 'diagnosis' of student performance in OSCEs that informs remediation has often taken place in the absence of real-life clinical performance with patients. (Hauer et al. 2008). Most recently, a major thematic review of the literature around remediation concluded that there still remains an absence of research exploring whether remediation in this poorly performing group provides a long lasting effect in improving academic performance (Hauer et al. 2009)

Therefore, it remains unclear whether we should review current methods of retraining and repeat assessment for this group. We set out to investigate this using longitudinal monitoring of performance of cohorts of undergraduate students within a single medical school.

Methodology

Within our own institution's MBChB programme, students' clinical performance is assessed using high stakes OSCE at the end of years 3, 4 and in the final year. Each OSCE follows a diet of clinical placements and other teaching, supported by in course and workplace based assessment. Each OSCE is constructed with 16-20 stations within an overall testing time of approximately 3 hours. Standard setting is undertaken using the borderline regression method and standard error of measurement which has been described previously (Kramer et al. 2003; Pell et al. 2006; Hays et al. 2008). Typical reliability coefficients for our OSCEs are in the range of 0.7-0.8 as measured by Cronbach's alpha.

A university-wide grading system is applied to the numerical distribution of OSCE marks in order to convert them to an A-F scale (A=Excellent, B=Good pass, C=Clear Pass, D=Borderline, E=Clear Fail, F=Bad Fail). The borderline grade is a narrow range with the lower point defined by the aggregate passing mark plus 1 standard error of measurement. This permits student achievement to be recorded centrally by the University, and allows the comparison of performance across years of the course. OSCEs are complemented with written, knowledge based papers with a similar grading system, using the Ebel method for standard setting (Cusimano 1996). Students achieving passing requirements for each assessment are allowed to progress, whereas those failing either component are subject to remediation and retest.

Failing students in years 3 and 4 of our course are typically offered a period of remediation followed by a resit within the academic year. Remediation usually involves an initial ‘diagnostic’ phase, where students are interviewed after assessment failure, feedback is given, academic and non-academic problems explored and a remediation plan agreed. Remediation takes place in a targeted clinical placement, incorporating feedback of performance post-OSCE, using experienced clinical supervisors to monitor performance. This is typical of many of the current reported models of remediation (Hauer et al. 2008; White et al. 2009). This remediation takes place over a 6-8 week period, followed by a resit OSCE of comparable rigour to the main assessment. This assessment is constructed from OSCE stations used in main diets, with previously determined passing scores and acceptable station level metrics. In the case of very poorly performing students or finals failure, a full year repeat of study is mandatory, with accompanying student support and extended remediation.

In this retrospective study, we reviewed university assessment records of student performance in OSCEs across a five year period (2004-2008). We identified all those students who have achieved a borderline grade or fail in at least one of their OSCE assessments for in years 3, 4 or 5. For this set of students, we have then extracted other OSCE performance data to build a picture of individual, longitudinal student performance for this specific group. For students to be included in the analysis, they must have undertaken a minimum of two OSCE assessments, not including resits.

The data used for the analysis is routinely generated from our OSCE assessments (including routine post-hoc analysis as a result of borderline regression methodology) and student performance record. Data was analysed in long format (i.e. 1 row per individual assessment), which means that when the arithmetic means are compared, some students appear up to three times, and some assessors will have assessed a number of students. However, in the limited ANOVA analysis, test statistics are not close to the critical 5% value, and the assessments are separated by a year in time, with very minimal effect of hierarchical clustering. Resits have *not* been included in the analysis as these could not be regarded as independent measures.

Because the subset of our data of most interest is a biased subset of the main data (i.e. the underperforming group), we have sought not to over-interpret this data, and kept statistical analyses to a minimum. Where we have quoted p-values, this has been done to help understanding of differences rather than state categorically that this is the probability of a particular difference occurring by chance in our non-random data. Furthermore, we anticipate that this simple approach is better understood by colleagues, can be applied to their own data sets, and will stimulate discussion on this important area of assessment.

Underperformers were then categorised, dependant on their pattern of performance and compared longitudinally with the mean performance of each year group at each stage of the MBChB programme

Results

Across the five year period analysed, we reviewed 13 whole OSCEs, reflecting the performance of approximately 3200 student assessments. We excluded year 5 results from the first year of the study, and year 3 at end of the study, as this paper deals with longitudinal performance and progression. Within this total population, 230 students received at least one borderline and/or fail grade from 2 or more OSCE assessments. For 105 of these students, we only have 2 years of assessments (dependant on the cohort, intercalation or departure from the course). This leaves 125 students for whom we have 3 years of OSCE assessment data and who received at least one borderline and/or fail grade. Table 1 gives a summary of the performance of these 125 students, and demonstrates that a student attaining a single fail grade but with otherwise good performance in the other years is extremely unusual.

TABLE 1 HERE

Three major profiles may be identified from our data, which account for the performance of approximately 60% of the underperformers.

- No failures but at least 2 borderline grades (26.4%)
- Single failure & at least 1 borderline grade (24%)
- Multiple failures (7.2%)

Table 1 demonstrates that 60.8% of students obtained at least one borderline grade with no failures, and approximately half of these obtained more than one borderline grade. A lower proportion of students (32%) fail a single OSCE and have at least 1 other borderline grade, of whom 5.6% have obtained two borderline grades.

Within the group of 125 students, 7.2% have multiple OSCE failures. Almost 70% of poorly performing students will show a repeated pattern of poor performance, with approximately two thirds of these (60%) failing at least one OSCE. A single OSCE fail with no evidence of other fail/borderline performance was highly unusual in the analysis, reflecting only 8% of students.

Longitudinal analysis of performance data

Table 2 compares the performance of the 125 underperforming students at year 3 with their grade at year 5, and it is clear that the performance in this group declines with progression across the course, despite episodes of remediation and resit for many of these students.

TABLE 2 HERE

Returning to the full subgroup of 230 underperforming students, a similar pattern emerges. By converting the individual student grade to a numerical scale (A=5, F=0), and comparing the mean grade over the three years, we find a significant difference in year group means ($p < 0.001$). 43.7% of year 3 students attain A, B or C grades, compared with 19.8% and 31.9% respectively for years 4 and 5 (although it should be noted that a small proportion of serially underperforming students leave before years 4 and 5).

The mean profile demonstrated by the poorly performing subgroup of 125 students is replicated by the entire body of students in the dataset; with the difference between the mean grade in years 3 and 4 being about one third of a grade, with some recovery seen however in final year students, highlighted in table 3. A comparison of means using ANOVA shows there is a significant difference between means (df =2, F= 41.7, P<0.001), with the Bonferoni correction the post hoc comparisons give the difference between each pairs of means as significant at the 5% level (Field 2009).

TABLE 3 HERE

From Figure 1 it can be seen that there is some similarity between the mean grade profiles of the entire cohort and the poorly performing subgroup. However, it should be noted that each year there an increase in the difference between these two groups.

FIGURE 1 HERE

Discussion

Despite the weight of literature focussing on high-stakes criterion based assessment, little exists to inform of the future progression of failing undergraduate medical students. Work has focused on short term impacts within a model of test-remediate-retest, demonstrating that the majority of students will pass resit OSCE assessments after a period of remediation, in keeping with published data (Sayer et al. 2002; Pell et al. 2009; Cleland et al. 2010). However, this retrospective study has shown that existing short remediation programmes offered at our school do not achieve the longer term goal of sustained improvement in future OSCE assessments for the majority of poorly performing students. Longitudinal review of these students' OSCE profiles reveals that the majority of candidates failing an OSCE assessment have an additional performance of borderline or fail. These findings begin to deal with need for longer term performance data to help us look critically at the remediation and further assessment of underperforming students, at least within an OSCE context (Hauer et al. 2009).

Why then does this research suggest that these underperformers pass resits, but gain little lasting benefit? Whilst resits will be of similar rigour and standard to those OSCEs undertaken by the whole student cohort, we must consider the environment in which the resit occurs. Whilst multiple models of remediation are described, they show commonality in both faculty-centred and learner-centred behaviours, with both interventions likely to be short term. Remediation is likely to take place with additional student support, often in smaller groups and with few or no additional distractions from other assessments or other course requirements. Students (and supervisors) efforts are predominantly focused on passing the resit assessments, and remediation programmes are likely to be tailored to improve performance to the required competency to pass.

Of more concern is that many poorly performing students in our study deteriorate between years 3 and 5, even though the very worst may leave the course (and hence do not form part of our study). What might be the reasons for this? There is clear evidence from other research and our annual detailed analysis of the OSCE data that examiners often find difficulty in agreement on student performance at intermediate levels within programmes, and this is revealed in higher levels of non-student variance which can prove resistant to interventions to improve station quality (Pell et al 2009; Pell et al 2010).

We believe that this variance is likely to work to the advantage of underperformers, perhaps as a result of assessor perceptions that students are only mid-way through the course, have scope to improve and are not undertaking higher-order skills and behaviours. This interpretation may in fact falsely reassure faculty and students about levels of ability, whilst preventing opportunities to identify and support poorly performing students.

The drop in mean OSCE grade between years 3 and 4, and the recovery in grade for the whole student cohort by year 5 suggests that other factors are impacting on this process. These may include the introduction of specialist outcomes and an associated expectation of higher levels of performance (in our own programme, clinical placements in Year 4 are specialist (e.g. psychiatry and paediatrics)), building on previous ‘general’ clinical experience. Our own experience locally suggests that there is better agreement on the minimally competent student within the Final year, with much lower levels of between group variance (Pell et al 2009). Similarly, these rising expectations of performance may be coupled, for some students, with a lack of ability to deal with the competing demands of OSCE preparation and required standard of in-course work, in an environment quite different to that in which tailored remediation for underperforming students takes place across the majority of institutions.

Although this study is based on significant numbers of students and ‘real life’ data (rather than the control of an interventional study), it is not without limitations. This data is drawn from only one medical school programme, and would be strengthened by collaborative approaches in other institutions, using similar methodologies. During the 5 year period of study, continued refinement of our OSCE programme has continued, with interventions to improve station level quality that have been particularly successful in year 4 and 5 OSCEs, and this may have the effect of further highlighting student performance issues that were once hidden within assessment quality concerns. The very recent introduction of year-specific OSCE assessor training may bring about a reduction in error variance (i.e. between group variance and/or poor assessor agreement) previously described for year 3, and we are currently examining to see if this is associated with an increase in underperformance.

Although there is considerable variation in the design of undergraduate medical degree programmes, it is likely that our findings are more widely generalisable in relation to serial underperformance. These results pose a number of questions for institutions in relation to identifying and supporting this group of underperforming students.

Whilst this research is in its early stages, further analysis of routinely collected OSCE data should allow us to more effectively profile students, and attempt to anticipate problems, especially given that 65% of our underperformers fall into one of three performance profiles. Using this data to predict and identify students at risk of serial failure should not be punitive but a supportive process. Effective remediation models need to be explored that provide *longitudinal* involvement by both Faculty and students, clearly measurable outcomes and allow effective longer term monitoring for this group of students. Similarly, the nature, scope and standards of resit assessments should be reviewed in context of this study's findings.

We are currently undertaking further research to understand more about the characteristics of the underperforming group which may assist in more tailored support. We would echo the calls of Hauer et al in the recommendation of collaborative programmes of research that examine both models of remediation and alternatives to the traditional assessment methods of retest.

Conclusion

Models of assessment where OSCE failure leads to directed remediation followed by resit, are not the optimum for ensuring good levels of performance in weaker students. Longitudinal performance profiles for students with fails or borderline passes suggest a failure to improve academically, with high rates of further failure or borderline behaviour. The nature of 'traditional' remediation programmes, whilst successful in the short term, may generate superficial learning with little lasting effect. Collaborative research to explore alternative models of remediation and resit policies, coupled with longitudinal data on student performance need/should be undertaken. Weak students need additional time to consolidate existing learning, and alternative models of remediation and sequential testing may prove attractive when coupled with methodologies to track the impact on longitudinal performance.

Acknowledgements: We are grateful to Dr PJR Harkin for his assistance in the preparation of the dataset.

References

Cleland J, Mackenzie RK, Ross S, Sinclair HK, Lee AJ. 2010. A remedial intervention linked to a formative assessment is effective in terms of improving student performance in subsequent degree examinations. Med Teach 32:e185-e190

Cizek GJ, Bunch MB, editors. 2007. Standard Setting. California; Sage

Cohen DS, Colliver JA, Robbs RS, Swartz MH. 1997. A Large-Scale Study of the Reliabilities of Checklist Scores and Ratings of Interpersonal and Communication Skills Evaluated on a Standardized-Patient Examination. Advances in Health Sciences Education 1:209-213

Cusimano MD, Cohen R, Tucker W, Murnaghan J, Kodama R, Reznick R. 1994. A comparative analysis of the costs of administration of an OSCE (objective structured clinical examination). Academic Medicine 69(7):571-6.

Cusimano M. 1996. Standard setting in Medical Education. Academic Medicine. 71(10):112-120.

Field A . 2009. Discovering statistics using SPSS. 3rd edition. London; Sage 55-56

Hays R, Sen Gupta T, Veitch J. 2008. The practical value of the standard error of measurement in borderline pass/fail decisions. Medical Educ 42:810-815

Hauer KE, Teherani A, Irby DM, Kerr KM, O’Sullivan PS. 2008. Approaches to medical student remediation after a comprehensive skills examination. Med Educ 42:104-112

Hauer KE, Ciccone A, Henzel TR, Katsufrakis P, Miller SH, Norcross WA, Papdakis MA, Irby DM. 2009. Remediation of the Deficiencies of Physicians Across the Continuum from Medical School to Practice. A Thematic Review of the Literature. Academic Medicine 84(12):1822-1832

Kramer A, Muijtjens A, Koos J, Dusman H, Tan L, van der Vleuten C. 2003. Comparison of a rational and an empirical standard setting procedure for an OSCE. Medical Educ 37(2):132–139

Pell G, Roberts TE. 2006. Setting standards for student assessment. International Journal of Research & Method in Education 29(1):91-103

Pell G, Boursicot K, Roberts T. 2009. The trouble with resits…. Assessment & Evaluation in Higher Education 34(2):243-251

Pell G, Fuller R, Roberts T, Homer M. 2009. Comments on within-station between-sites variation. Medical Educ 43(10):1021-1022

Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: a review of metrics. Medical Teacher. in press

Martin IG, Jolly B. 2002. Predictive validity and estimated cut score of an objective structured clinical examination (OSCE) used as an assessment of clinical skills at the end of the first clinical year. Med Educ 36:418-425

Regehr G, MacRae H, Reznick RK, Szalay D. 1998. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Academic Medicine 73(9):993-997

Ricketts C. 2010. A new look at resits: are they simply a second chance? Assessment **&** Evaluation in Higher Education **35**: 351 – 356

Roberts C, Newble D, Jolly B, Reed M, Hampton K. 2006. Assuring the quality of high-stakes undergraduate assessments of clinical competence. Medical Teacher 28(6):535-543

Sayer M, Chaput de Saintonge M, Evans D, Wood D. 2002. Support for students with academic difficulties. Med Educ 36:643-650

Streiner DL, Norman GR, editors. 2008. Health Measurement Scales (4th edition). Oxford; Oxford University Press

White CB, Ross PT, Gruppen LD. 2009. Remediating Students’ Failed OSCE Performances at One School: The Effects of Self-Assessment, Reflection and Feedback. Academic Medicine 84(5):651-654

Tables

No. of Failed OSCE Assessments	Number of Borderline Grades	Number of Students	% of students	
No Fails	1	43	34.4	60.8
	2	30	24.0	
	3	3	2.4	
One Fail	0	10	8.0	32.0
	1	23	18.4	
	2	7	5.6	
Two Fails	0	8	6.4	7.2
	1	1	0.8	
Total		**125**	**100**	**100**

Table 1: Summary of the performance of poorly performing students during OSCE assessment.

	Performance Improved	**Performance the Same**	**Performance Deteriorated**	**Total**
Number of Students	17	35	73	**125**
% of students	13.6	28.0	58.4	**100**

Table 2: OSCE Performance at year 3 compared to grade at year 5 – underperforming group

Year	Entire student population		Underperforming sub-group	
	Number of Students	Mean OSCE Grade	Number of Students	Mean OSCE Grade
3	1689	3.45	125	2.86
4	967	3.10	125	2.02
5	1343	3.24	125	2.06

Table 3: Mean OSCE performance - entire student population and underperforming sub-group

Figures

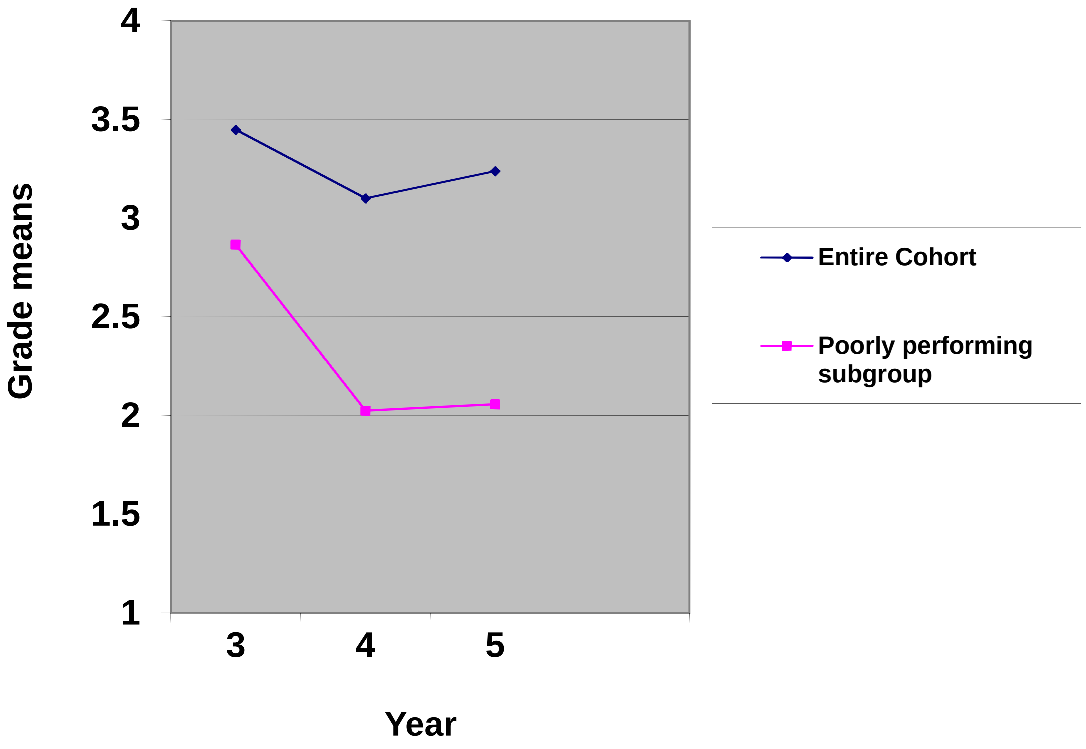

Figure 1: Graph of mean grades of students by year of programme (entire cohort and poorly performing subgroup)