Modelling the use of confidence intervals with the borderline regression method for final year undergraduate OSCE at the University of Southampton.

UNIVERSITY OF
Southampton
School of Medicine

McManus,B.N; Carr,N.J; Anderson,F.H.; Holloway,J.A.; Field,J.H; & Rushworth,S.M.

McManus,B.N, BM finals OSCE coordinator, Faculty of Medicine, University of Southampton, Southampton General Hospital MP801, Tremona Road, Southampton.

Abstract:

Background and Purpose
Medical students undertake Objective Structured Clinical Examination (OSCE) as part of their BM finals, and as a high stakes examination,, improvements in the reliability and validity of these exams is desirable. We wished to model and pilot a novel use of the confidence interval (CI) and standard error of the measurement (SEM) with the borderline regression method, in line with recommendations by PMETB/GMC,[1,2] and in place of simple examiner global judgements.

Methodology
Students must satisfy two criteria to pass the BM finals OSCE: aggregate score and minimum number of stations passed. The SEM has been equated with CI[1] and applied to aggregate score[3-5]. We wished to introduce it into our examination, and also proposed a novel strategy to calculate the CI in the cut score for a single station. Using the standard error of the intercept and gradient we calculated the CI for these values, and used them in the regression equation to interpolate a new value of y when x is constant. We modelled these techniques to maximise the sensitivity and specificity of both criteria.

Results
In a cohort of 242 students, 6 failed >3 stations on global judgement. For 2 of them the mean grade was also below the threshold but none failed this criterion alone. Introducing borderline regression without adjustment, 23 students failed >3 stations but none on aggregate score. Recalculating the aggregate pass mark as mean cut score plus 1.96xSEM (upper 95%CI) considerably improved the sensitivity of the aggregate score criterion, which 6 students now failed.
For individual stations, using the gradient and intercept minus 1.96 x the respective Std Error for these constants (lower 95% CI) provided an adjusted cut score for each and considerably improved the specificity of this criterion. Students failed if their actual scores were below the adjusted cut score for >3 stations. 7 failed on this criterion.
Considering both criteria 8 failed the OSCE, 5 of whom failed both criteria. Observed agreement with global assessments rose from 92.1% to 98.35% (Kappa 0.32 to 0.71).

Discussion and Conclusions
The adjusted cut scores showed improved sensitivity and specificity for both criteria and improved agreement with global judgements. It was perceived to be fair to students, affording them the benefit of the doubt when considering individual stations, but protecting patient safety when decisions could be reliably based on 16 assessments. Since most students who failed did so on both criteria, the method was perceived to be more robust. The authors plan to pilot this new methodology, providing improved sensitivity, specificity and robustness on another cohort of students, before considering incorporating into the exam regulations.

1) Background:

- OSCEs have been used as a core component for several years at the University of Southampton. BM Finals uses a 16 station OSCE.
- Examiners for each station awarded a grade to indicate global assessment of student performance A—F, where grades E & F represent a fail for that station.
- A standard for overall OSCE set as a mean grade of D or above and not more than 3 station fails—to ensure minimum competence for passing students both overall and across the breadth of the curriculum Students are required to achieve minimum standard over >75% of stations.
- The problem here is that while relatively simple to administer this system is highly reliant on single examiners making very high stakes decisions accurately, and the system therefore lacks robustness.
- We have piloted and modelled a well described alternative method known as the borderline regression method (BRM)[6,-9] to standard set the OSCE component of finals. "Borderline" was defined as half way between grade "D" (equivalent to a borderline pass) and a grade "E" (equivalent to a borderline fail).
- We wished to maintain the minimum no of stations passed criteria in addition to a minimum overall performance.
- Cut scores were then calculated for all 16 stations using BRM and the students' actual scores for each station were compared to these cut scores for each station to arrive at a pass / fail decision for every student on every station. As previously students would be allowed a maximum of 3 out of 16 individual station fails ***(Criterion 1)***.
- Secondly the mean cut score for all 16 stations was compared to each students mean aggregated score. The students mean aggregated score had to exceed the mean cut score in order to pass the OSCE ***(Criterion 2).***

2) Results of initial pilot:

Overall OSCE outcome		Borderline system		Total	Borderline system pass rate %
		Pass	Fail		
Current System	Pass	218	18	236	
	Fail	1	5	6	
Total		219	23	242	90.50%
	Current system pass rate %			97.52%	

% Observed Agreement	92.14 %	No. who fail *(Criterion 1)*	23
% Chance Agreement	88.49 %	No. who fail *(Criterion 2)*	0
Kappa Coefficient of agreement	0.318		

Table 1

3) Problems:

Table 1 shows:

- Poor agreement between current and BRM system
- Significant rise in fail rate
- Therefore standards not equivalent—difficult to justify as:
  - There is evidence to suggest that checklists are not intrinsically more reliable than global ratings[3,6, 10-12.]
  - Our pilot showed no significant difference in reliability as measured using Cronbach's alpha between the current and the BRM system
    - (0.739 - current system vs. 0.732 - BRM)
- All students who fail do so on Criterion 1—minimum no of station passed
  - Suggests criterion lacks specificity as a screening tool for competence
- No students fail to achieve overall score
  - Suggests criterion lacks sensitivity as a screening tool for competence

4) Possible Solutions

Option 1:
Simply increase the max no of station fails

- Would decrease fail rate potentially to an equivalent to current system

But:

- Increases scope for compensation across stations
- No logical method described to standard set what the max no of station fails should be
- Doesn't address the lack of sensitivity of the overall criteria *(criterion 2)*

Option 2:
Acknowledge inevitable statistical error and incorporate confidence intervals into the standard setting:
1) Give the student the benefit of the doubt when decisions are made on a single station, thus failing students who have failed "beyond reasonable doubt" Thus deflate the cut score to a lower confidence interval.
2) Give patient safety the benefit of the doubt when decisions are more reliably based upon 16 different stations. Thus inflate the overall passing score to an upper confidence interval.

It was proposed these changes might

- Improve equivalence with current system by
  - Improving agreement with examiner judgements
  - Bring the fail rate more in line with examiner judgement
- Improve specificity of this criterion 1 without allowing more compensation
- Improve sensitivity of the overall performance criterion *(criterion 2)*

However to test this hypothesis the same data set was remodelled using option 2

5) Method:
Standard Error of Measurement:

- Use of the Standard error of the measurement well described to allow a CI around the overall score[1,3-5]
- Therefore easily applied to criterion 2
- Uses following formula (incorporating the reliability statistics)

$$s_E = s_x \sqrt{(r-1)}$$

Where s_E = Standard error of the measurement
Sx = Standard deviation
r = Reliability coefficient (Cronbach's alpha)

- For criterion 2:
  passing score = (1.96 x SEM) + mean unadjusted cut BRM cut score
- This represents the upper 95% CI

However not possible to calculate Cronbach's alpha for a single station thus not possible to calculate SEM for a single station.

Therefore a different approach is needed at single station level

6) Method:
Confidence intervals for the cut score on a single station:

- Unable to use SEM on a single station as unable to calculate a reliability coefficient.
- Can calculate the standard error for the intercept and the gradient for the regression equation.

  $y = gx + i$

  Where g = gradient and i= intercept
- 95% CI for each of these constants can be calculated as 1.96 x Std Error of constant
- Lower CI for each calculated as:
  = (g or i) - (1.96 x St err (for g or i))
- Allows a new regression equation to be computed and drawn as per Figure 1 (Black regression line)
- X in the equation remains = 2.5 (as before)
- A new adjusted lower CI cut score can be interpolated from this new equation (see figure 1 blue line)
- Student actual scores now compared to this adjusted cut score to determine pas / fail for each station.
- No of station fails counted for each student criterion 1 failed if > 3 of 16 (as before)
- Students still need to pass Criterion 1 & 2 in order to pass the OSCE.
- Results demonstrated in Table 2

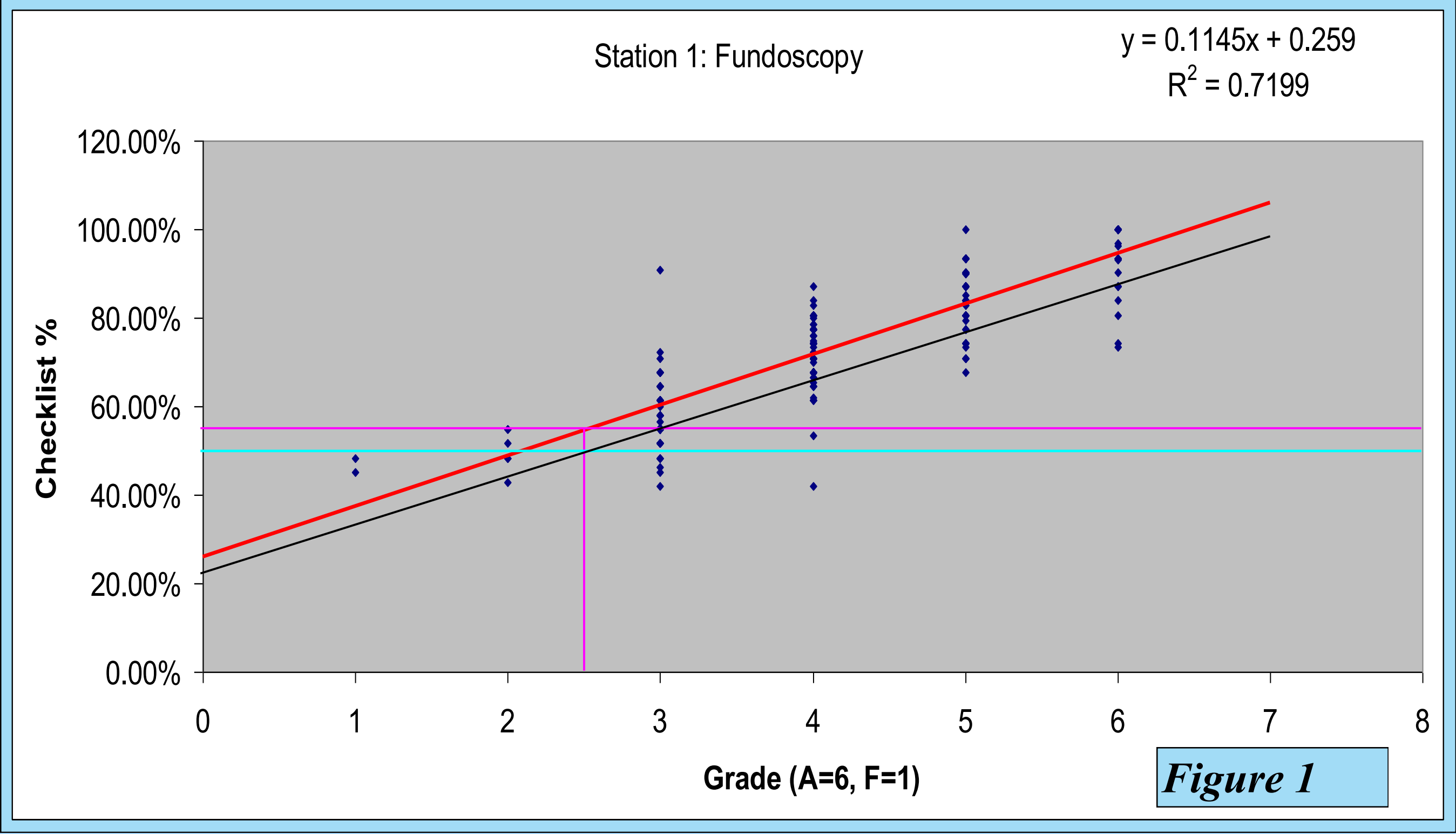

Figure 1

7) Results of re-modelling:

Overall OSCE outcome BRM		Borderline system		Total	Borderline system pass rate %
		Pass	Fail		
Current System	Pass	233	3	236	
	Fail	1	5	6	
Total		234	8	242	96.69%
	Current system pass rate %			97.52%	

% Observed Agreement	98.35 %	No. who fail on *Criterion 1*	7
% Chance Agreement	94.38 %	No. who fail on *Criterion 2*	6
Kappa Coefficient of agreement	0.71	No. who fail on **both** criteria	5

Table 2

8) Discussion & Conclusions:

Table 2 shows:

- Significant improvement in both agreement & thus similar fail rate to current system
- The system with similar reliability gives equivalent results to the current system
- The Specificity of Criterion 1 appears to have improved
- The Sensitivity of Criterion 2 appears to have improved
- Most students who fail now fail on both criteria making the system more robust against appeal
- Students have the benefit of the doubt when considering single stations
- Students who have beyond reasonable doubt failed to achieve minimal competence across >75% of the breadth of the curriculum fail. This high standard preventing compensation is maintained.
- Patient safety has the benefit of the doubt when we have the reliability of 16 stations. Students must therefore demonstrate beyond reasonable doubt that they are fit to practice.
- In both criteria students within the statistical error (corresponding to 95% confidence) either side of the relevant pass marks are treated similarly thus conforming with GMC guidance on assessment.[1,2]
- The model is a relatively straightforward extension of the borderline regression method and feasible within common statistical and spreadsheet software packages
- Use of the standard error of the measurement has been previously described applied to aggregate score[3-5] but incorporating confidence intervals at single station level to a criteria to prevent compensation between stations is novel so far as the authors are aware.
- This is a work in progress and will be remodelled on a new cohort in this summer's final examinations with a view to adoption as the formal standard setting procedure at Southampton

References

1. Postgraduate Medical Education and Training Board. "Developing and maintaining an assessment system - a PMETB guide to good practice." PMETB 2007.
2. General Medical Council. "Assessment in undergraduate medical education - Advice supplementary to Tomorrow's Doctors (2009)." GMC 2010.
3. Dauphinee, W.D., Blackmore, D.E., et al. "Using the Judgments of Physician Examiners in setting the Standards for a National Multi-center High Stakes OSCE." *Advances in Health Sciences Education* 1997; **2:** 201–211.
4. Smee, S.M., Blackmore D.E. "Setting standards for an objective structured clinical examination: the borderline group method gains ground on Angoff ." *Med Educ* 2001; **35:** 1009-1010.
5. Kilminster, S., Roberts, T. Standard Setting for OSCEs: Trial of Borderline Approach. *Advances in Health Sciences Education* 2004; **9:** 201–2097.
6. Boursicot KAM, Roberts TE, & Burdick WP, "Structured assessments of clinical competence" Association for the study of medicine: ASME 2007
7. Kramer, A., Muitjens, A., Jansen, K., Dusman, H., Tan, L. & van der Vleuten, C. "Comparison of a rational and an empirical standard setting procedure for an OSCE." *Med Educ* 2003, 37:132-139.
8. Wood, T., Humphrey-Murto, S., & Norman, G. "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* 2006; **11:** 115-122.
9. Boursicot KAM, Roberts TE & Pell G, "*Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools*." *Med Educ* 2007: **41**: 1024–1031
10. Regehr G., Macrae, H., Reznik, RK., & Szalay, D., "Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination." *Academic Medicine*. 1998; **73:** 993-997.
11. Hodges B., McNaughton, N., Regehr G., Tiberius, R & Hanson M., "The Challenge of creating new OSCE measures to capture the Characteristics of expertise" *Med Educ* 2002: **36:** 742-748.
12. Cohen, R., Rothman, Al., Poldre, P., & Ross, J., "Validity and generalisability of global ratings in an objective structured clinical examination." *Academic Medicine*. **66:** 545-548.