The impact of the inclusion of simulated patient ratings on the reliability of OSCE assessments under the borderline regression method

Short title: Including SP ratings in OSCE standard setting

Authors: Matthew Homer and Godfrey Pell

Acknowledgment: The authors we like to acknowledge the advice and encouragement of Dr. Richard Fuller, Academic Sub-dean, School of Medicine, University of Leeds.

Institution: Medical Education Unit, University of Leeds

Corresponding author: Dr Matthew Homer, Medical Education Unit, School of Medicine, University of Leeds, LS2 9JT. Tel: + 44 (0)113 343 4654; Fax, +44 (0)113 343 5260; Email: m.s.homer@leeds.ac.uk

Abstract

Background

In Objective Structured Clinical Examinations (OSCEs), the use of simulated patients (SPs) at many stations is a key aspect of the assessment. Often the SPs are asked to provide formal feedback (ratings) of their experience with the students under examination.

Aims

This study analyses whether and how exactly SP data can be best used to enhance the robustness of the formal standard setting process.

Methods

A retrospective statistical investigation into the relationship between SP ratings and those provided by the clinical assessors (criterion-based checklist scores and overall grades for each station) is presented. In addition, the paper also includes a study into the impact of the inclusion of the SP ratings in the formal standard setting process for OSCEs, particularly when pass marks are calculated using the borderline regression method.

Results

The general results of the analysis, including discussion of two distinct methods for the combining of the SP ratings and assessor judgements, are presented, and demonstrate that the inclusion of this additional data can have important effects on individual student results.

Conclusion

It is possible for the overall quality of the OSCE assessment process to be improved, with increased reliability, by combining assessor checklist scores and SP ratings.

Introduction

The borderline regression method

Under the borderline regression method, assessors produce individual marks against an item checklist that are accumulated as students meet particular criteria as the assessment at each OSCE station proceeds. See Newble (2004) for more background on the use of OSCEs in clinical assessments, and for the initial developments of borderline methods of standard setting see, for example, Livingston and Zieky (1982).

Kramer *et al.* (2003) and Wood *et al.* (2004) discuss comparisons of the borderline regression method with other methods of standard setting for OSCEs and generally conclude that this is a robust method for standard setting in OSCEs, with high levels of reliability in comparison to some other methods. De Champlain *et al.* (2001) carry out a weighted statistical modelling exercise in order to maximise the accuracy with which mastery level can be estimated within the OSCE context, but their work does not include the use of simulated patient ratings.

Typically the maximum assessor mark (i.e. checklist score) is of the order of 20-30 for a particular station. Separately, the assessors also provide an overall (global) grade for the station – for the particular OSCE arrangements discussed in this paper (Year 4[1]), these judgements range from "clear fail" to "excellent pass" – coded on a scale from 0 to 4 with 1 as the key "borderline" grade – see Table 1 below:

TABLE 1 HERE

Each individual station pass mark is then calculated by regressing the set of station assessor checklist scores on the corresponding global grades given by the assessors. The pass mark for the station is given by the checklist score on the regression line corresponding to the borderline grade as demonstrated in Figure 1.

FIGURE 1 HERE

[1] That is, the fourth year (out of five) of an undergraduate medical degree (MBChB).

It is important to note that it is possible for a student to be awarded a "fail" grade but still pass the station having achieved a sufficiently high checklist score, and similarly, to be awarded a "pass" grade but fail the station based on low assessor checklist score.

The overall OSCE pass mark is the sum of the individual station pass marks, adjusted upwards using the Standard Error of Measurement (SEM). Streiner and Norman (2003, p.142) give more details on the adjustment of passing scores using the SEM.

Simulated patient ratings and this paper

At certain stations where it is deemed pedagogically appropriate in terms of the aims of the assessment, simulated patients (SPs) are also asked to rate the students' "professional manner". There are two main reasons for asking SPs to rate the perceived quality of the treatment that they receive: firstly, to emphasise to all stakeholders the importance that is placed on valuing patient-student interactions, and, secondly, to provide an additional motivation for SPs in taking their role seriously and to encourage them to act with due care and attention in participating in the OSCE.

Table 2 shows details of the SP ratings scale:

TABLE 2 HERE

Having afforded the SPs the opportunity to rate the students' professional performance, it is natural to ask whether (and if so, how) this information might be usefully employed to enhance the robustness of the OSCE assessment process.

This study will therefore consider two models for the formal inclusion of the SP ratings in the pass mark calculations:

1. Adding assessor checklist scores and SP ratings together to produce a combined "mark" (this will be referred to throughout as **method 1**)
2. Adding assessor global grades to SP ratings to produce a combined "grade" (**method 2**).

The analysis will investigate the effect that the inclusion of the SP ratings might have on the pass rates, both overall and at the station level, and on the apparent reliability and overall quality of the assessment process. There will also be discussion of alternative approaches that do not use the SP ratings as an integral part of the calculation, but rather treat them as a complimentary aspect of the OCSE assessment process. In the literature there are examples of studies where the quality of SP ratings have been compared to those of clinical experts, sometimes quite favourably; see, for example, McLaughlin *et al.* (2006) who found that SP ratings and feedback were generally well-respected by (third year) students undergoing an OSCE. However, there is little evidence in the literature of studies where SP ratings have been utilised formally in the standard setting, certainly where the borderline regression method is employed in calculating the passing scores.

Two successive years of OSCE data, Year 4 examinations from 2006 and 2007, will be utilised in this retrospective study, and Table 3 gives an overview of this data.

TABLE 3 HERE

Methods

Background analyses

Before attempting to utilise the SP ratings in a revised form of pass mark calculation, it is important to carry out an exploratory analysis get a sense of the degree to which the SP ratings show an association with both the assessor checklist scores and the assessor grades. If there were no association or evidence of an association in the "wrong" direction then this would raise serious concerns about the appropriateness of formally combining the SP ratings with the checklist scores/grades of the assessors. For completeness, a brief exploration of the relationship between the two "marks" awarded by the assessor (checklist score and overall grade) will also be carried out.

Comparing SP ratings to assessor checklist scores

A correlation-based analysis was carried out for each station across the two years of data. In summary:

- There was always a significant positive Pearson correlation between the two variables ($p<0.05$). The effect size (R-squared) was, however, sometimes quite small with a lowest value across the 17 stations being 0.02, implying that only 2% of the variation in assessor checklist scores was explained by the variation in SP ratings. The highest R-squared value was 0.29 with most values of the order of 0.1 or less. Scatter diagrams indicated that there was usually a wide range of SP ratings for the same assessor checklist score and vice versa.

Comparing SP ratings to assessor grades

The mean SP ratings were higher than the assessor grades across all 17 stations, since the two scales are not directly comparable, despite sharing the same range of

numerical values (compare tables 1 and 2). The general pattern in the relationship between assessor and simulated patient ratings across the set of stations can be summarised as follows:

- At each station, the two sets of "grades" always showed a significant positive correlation ($p<0.05$), varying in size from 0.300 to 0.563[2]. As was the case with the checklist scores and SP ratings, there was generally a wide spread of assessor grades for the same SP rating, and vice versa.

Comparing assessor checklist scores to assessor grades

For each of the 17 stations where SP data was available, the Pearson correlation coefficient between assessor checklist scores and assessor global grades was calculated:

- At each station, the two variables always showed a significant positive correlation ($p<0.05$), varying in size from 0.659 to 0.865. As above, scatter graphs exhibited a spread of assessor checklist scores for the same assessor global grade, and vice versa.

The validity of combing SP ratings with assessor checklist scores or grades

It is clear, then, that generally there is a statistically significant positive relationship at the station level between both assessor checklist scores and SP ratings, and between assessor grades and SP ratings (whilst sometimes small in effect size). It seems therefore that combining either of these two pairs of assessment "scores", using a simple sum of the two, is methodologically justifiable since they pairs do show some

[2] Pearson's correlation was used here in spite of the ordinal nature of both of the grades, since this early analysis is exploratory only.

degree of positive association with each other. A comparison of the relative sizes of the correlations implies that there is less redundancy in combining the SP ratings with the assessor checklist scores, than with the assessor global grades. This observation will be commented on further later in the paper.

Measuring the efficacy of the assessments

In addition to the calculation of pass marks, statistical metrics are traditionally employed to enable judgements to be made as to how "well-behaved" the scores and grades at each station are, and to allow for comparisons between stations in terms of the quality of their "performance" as assessments. This allows for retrospective evaluation of the stations in terms of their fitness for purpose, and therefore allows for possible improvements to be made in stations that might be re-used in the future. For OSCEs, the station-level statistics typically produced are as follows:

- *Adjusted R-square* gives the proportion of the variance in the scores explained by the regression model taking account of the number of predictors in the model (in this case just one, assessor grade).
- The *inter-grade discrimination* is the slope of the regression line and indicates how many marks there are between, for example, a borderline pass and a good pass. Stations with low values of this statistic show little discrimination between students, and higher values, separating out differing student performance, are generally preferred.
- The *number of failures* shows the proportion of students that failed the particular station. In addition, overall OSCE pass rates are also produced.
- *Cronbach's alpha if the item (station) is deleted* indicates how well each station is measuring abilities similar to those measured at the other stations – if the value

is lower than the overall alpha value across the full set of stations then this is indeed the case, but if it is higher than this value then the station might be a cause for concern[3].

The impact of the inclusion in the calculations of the SP ratings on the quality of the OSCE assessment data will be assessed using these measures.

Results

A standard analysis was first carried out producing all the relevant pass rates and station diagnostics based only on the assessor checklist scores and assessor grades (ie without the inclusion of the SP ratings). This was then used as a baseline for assessing the effect of the inclusion of SP ratings in the calculations under the following two distinct methods.

Adding the SP ratings to the assessor checklist scores (method 1)

For the 17 stations with SP data, the SP ratings were added to the assessor checklist scores and then the borderline regression method was used to calculate the pass mark(s) for these new "marks" regressing on the assessor grades[4].

---

[3] Generalizablity theory is not employed in this study for measuring reliability because of limitations in the nature of the OSCE data available. Firstly, there is insufficient crossing of items (stations) with assessors, and secondly, the OSCE stations have checklist scores that vary in their maxima, and in their mean (expected) student score – under such conditions, g-coefficients are likely to be depressed in value, and do not necessarily give an accurate measure of reliability.

[4] The assessor marks are on a longer scale (generally of the order 0 to 30) and have typically higher numerical values compared to the SP ratings. Hence, the analysis presented here is relatively conservative in the sense that the weighting in the combined "mark" of the SP ratings

For the 2006 data, it was found that 13 students failed (out of a total of 264), one more than previously, with the other 12 students the same individuals who failed under the original calculations. For the 2007 data there were 18 failures in all (out of 267), with two additional students failing under the new method of calculation. Hence the inclusion of the SP ratings by adding them to the assessor checklist scores increased the failure rate by 0.38% and by 0.75% in 2006 and 2007 respectively. Both of these changes in the number of overall failures are non-significant according to the McNemar test.

Furthermore, the value of Cronbach's alpha for the assessment as a whole increased slightly (from 0.725 to 0.749, and from 0.759 to 0.777, for the 2006 and 2007 OSCEs respectively) indicating an increase in reliability amongst the set of items (i.e. stations) on inclusion of the SP ratings under method 1 for both years of data.

In comparing the remaining diagnostics it was found that:

- The *adjusted R-square* figures show an increase in this statistic for 16 of the 17 stations under investigation, generally of the order of a few percent and indicating a higher percentage of the variance in the checklist scores explained under this method compared to that not including SP ratings.

---

is comparatively low. A different weighting could be chosen that would give more emphasis to the SP ratings and this might be appropriate for future research.

- The *inter-grade discrimination* values (the slopes of the regression lines) are all greater than before but this might be as expected since the scale of measurement has been increased on adding the SP ratings to the assessor checklist scores, thereby allowing for a greater variation in the combined mark.
- Perhaps the most interesting aspect of these new results is the impact on the number of failures at the station level. Overall, whilst the vast majority of students are unaffected under the new method, there were 6 per cent more student failures at the station level (113 more in total out of 1848=264×7) in the seven 2006 stations compared to the analysis where SP data was not included. This is a significant increase in the mean number of individual station level failures per student across the entire assessment (from 2.69 to 3.12; paired sample t-test, t=7.49, df=263, p<0.001). For 2007, again most students were unaffected but this time there were 2.5% *less* student station failures overall (68 less in total out of 2670 = 276× 10) at the station level. Again this is a significant change in the mean number of station level failures, this time a decrease (from 3.27 to 3.02; paired sample t-test, t=6.63, df=266, p<0.001). These results suggest that adding SP ratings to assessor checklist scores has a varied but significant impact upon the number of station level student failures, with the variation dependent upon the complex interaction between the distribution of assessor checklist scores, assessor grades and SP ratings.
- The *Cronbach's alpha if the item is deleted* figures for each station have all increased, and this is consistent with the overall alpha increasing. More importantly, **all** these values for alpha at the station level are less than the respective overall alpha figures for the set of stations (0.749 in 2006, 0.777 in 2007). Under the original method of calculation this was not the case for three

stations of the total of 36 across the two years[5]. Hence, adding in the SP ratings under method 1 has led to a small but meaningful improvement in the Cronbach's alpha measures of the quality of the assessments.

Adding the SP ratings to the assessor grades (method 2)

For the stations with SP data, the SP ratings were added to the assessor grades and then the borderline regression method was used to calculate the pass mark(s) for the assessor checklist scores regressing on these new combined "grades". In this approach, the two original ratings/grades share the same scale of measurement and so adding them is equivalent to giving them equal weight. There are strong objections to (as well as some arguments in support of) these equal weightings but to allow for a complete comparison of methods of inclusion of SP ratings, it was thought important to report on the results of this approach.

However, there is an additional complication with implementing this method since the question as to what exactly is a borderline performance as far as simulated patients are concerned has to first be resolved. Recall (see Table 2) that a (coded) grade of 1 is *disagree*, 2 is *neutral* and 3 as *agree* as to the acceptability of the students' professional performance in the eyes of the patient. Hence it seems logical to state that a borderline performance as judged by an SP would be at 2. The combined borderline "grade" (assessor plus SP) at such stations would then be

3 = 1 (for assessor grade) + 2 (for SP ratings))[6]

[5] Note here that this statistic is relevant to the full set of stations, and not only to those where SPs were employed.

Accepting this argument it was found that, for the 2006 data, 14 individual students failed to reach the required total pass mark, 12 of whom had failed under the original method. Hence, the inclusion of the SP ratings by combining them with the assessor grades has led to a small increase in the number of overall student failures - 2 out of 264 students in total; that is a 0.75 per cent increase. For 2007, 20 students failed under this new method of pass mark calculation; an increase of 4 student failures (out of 267) compared to the original calculation, this time a 1.5 per cent increase in the number of student failures. For both years of data these increases are not significant (McNemar test).

- The *adjusted R-square* figures are all considerably lower than the original values (across the 17 stations with SP ratings) indicating lower proportions of variance explained in each of the station models compared to the original calculations.
- The *inter-grade discrimination* values (the slopes of the regression lines) are systematically lower since the horizontal scale (the range of the possible grades) has doubled on combining the assessor grades and the SP ratings. A fair comparison would therefore entail doubling the inter-grade discrimination values for new data – on doing so, six are lower and eleven higher across the pair of assessments showing, on balance, an improvement in the discrimination between the candidates under this method.

---

[6] The borderline grade at the stations that do no have SP ratings would remain as 1.

- As was the case when the assessor checklist scores and SP ratings were combined under method 1, it is the number of failures at the station level where the full effect of the new calculations on passes and failures can be seen. Whilst overall, the majority of students were unaffected by the new method, in 2006 there were approximately 13 per cent more student failures at the station level (236 more in total out of 1848=264×7, a significant increase in the mean number of station level failure, 2.69 to 3.58; paired sample t-test, t=14.87, df=263, $p<0.001$) compared to the original analysis where SP data was not included. In 2007, there were 2% more failures at the station level across the 10 stations with SP ratings (46 more out of a total of 2670=267×10) and again the mean number of student level failures for the whole assessment increased by a significant number (from 3.27 to 3.45; paired sample t-test, t=-7.08, df=266, $p<0.001$). Hence, for both years of data there has been a significant increase in the number of station level failures on adding assessor grades and SP ratings to produce a combined "grade".
- The *Cronbach's alpha if the item is deleted* figures would not change since the total (assessor) checklist scores have not changed.

Discussion and conclusion

The inclusion of SP ratings in standard setting for OSCEs

The analyses have demonstrated that the inclusion of the SP ratings (under methods 1 and 2 respectively) had only a small effect on the overall pass/failure rates for the OSCE assessment in question. In addition, the first method (assessor checklist scores plus SP ratings) did not show any significant negative impacts on the station-level diagnostics that would be a cause for concern about the use of SP ratings in OSCE

pass mark calculations. However, for the second method, where assessor grades were combined with SP ratings, the diagnostics were not generally as good. This is consistent with the earlier comments regarding the additional redundancy that exists between the SP ratings and the assessor global grades, in comparison with that of the former and the assessor checklist scores.

At the station level, the effects of the inclusion of SP ratings were more marked and more varied – the second method (adding SP ratings to assessor grades) created more than twice as many additional station-level failures (13%) than the first (adding SP ratings to the assessor grades, 6%). Thus whilst both methods would have an impact on certain students in terms failing, this would be much greater under method 2, partly as a consequence of the stronger weighting given to the SP ratings under this method.

The exploratory statistical analysis assessing the relationships between the SP ratings and the assessor checklist scores and grades indicated that the SP ratings appear to relate more closely with the assessor grades than they do with the assessor checklist scores. This is, perhaps, not a surprising result, since this latter pair are both overall judgements, whereas the assessor checklist scores are formed by a different process – that of a ticking off individual specific criteria on a checklist to produce a cumulative total. Indeed, perhaps there is an echo here of Cohen *et al.* (1996) and Regehr *et al* (1998) both of which conclude that global rating scales tend to have higher levels of both reliability and validity compared to checklist scores.

Overall, then, the picture is a complex one and the decision as to whether to include SP ratings into the formal assessments of medical students, and if so how to do so, remains a decision that rests on more than a purely technical analysis. However, whilst both methods appear to be statistically defensible, the large increase in station-level failures, and the deterioration in the diagnostics under method 2 both strongly suggest that the method 1 might be preferred, and would be likely to provide a less dramatic increase in student failures if adopted.

Broader considerations of the two methods

The philosophical and practical question as to which of the two methods is the most appropriate is open to debate. However, method 2 - combining grades with SP ratings - has the effect of giving equal importance to these two judgments; a decision that implies equating the judgement of medical professionals to that of non-specialists. There might be considerable opposition to the implementation of such a decision from many of the concerned stakeholders – the medical profession, the student body, and also, the general public who need to be re-assured that the procedures in place to certificate new entrants to the medical profession are sufficiently reliable and are based on the informed opinions of experts. In essence, it is likely to be felt that SP ratings do not carry sufficient face validity to be used in a manner that weights them so heavily. In addition, there are further questions raised with the validity of method 2 in that, whilst both assessor grades and SP ratings are global judgements of student performance, they are measuring very different attributes of that performance. However, method 1 could enjoy the same criticism, though arguably to a lesser extent since the SP ratings have a lower weighting under this method.

An isolated approach

There is one final point to be made – that there are other, alternative, methodologies available for utilising the SP ratings that would not directly affect the calculations carried out under the borderline regression method. One such approach would be to employ an entirely separate system where the ratings are considered by examiners in isolation from the assessor checklist scores and grades. Such an approach would indeed help in avoiding all the issues that have been raised in the earlier discussion concerning, for example, how exactly to include the ratings, and what appropriate weighting to give. However, any such system where the SP rating were intended to "really count" by providing an additional hurdle would again come up against the issue of the insufficiently high face validity of these ratings[7]. Furthermore, it should be remembered that the quality control of SPs (and subsequently of their ratings) is to some extent often beyond the influence of the institution carrying out the assessments, and therefore a separate system for formally employing these ratings brings with it its own serious objections.

[7] A number of medical schools currently use SP ratings as formative assessments, counselling those students who receive a relatively high number of poor ratings across the OSCE as a whole.

Practice points

- The use of simulated patients in OSCEs forms a key part of the assessment.
- SP ratings can be used to enhance the standard setting process.
- Under the borderline regression method, adding the SP ratings to the assessor checklist marks appears to be a robust way of incorporating such data, leading to the improved reliability of the assessment.

Notes on contributors

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.

Godfrey Pell is a senior statistician who has a strong background in management. Before joining the University of Leeds he was with the Centre for Higher Education Practice at the Open University. Current research includes standard setting for practical assessment in higher education, and the value of short term interventionist programmes in literacy.

References

Cohen, D.S., Colliver, J.A., Robbs, R.S. and Swartz, M.H. (1996). 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.

De Champlain, A.F., Margolis, M..J., Macmillan, M.K. and Klass, D.J. (2001). Standardized Patient Test: A Comparison of Case and Instrument Score-based Models Using Discriminant Function Analysis, *Advances in Health Sciences Education*, **6:** 151–158.

Kramer, W.M., Muijtjens, A., Jansen, K., Düsman, H., Tan, L. and van der Vleuten, C. (2003). Comparison of a rational and an empirical standard setting procedure for an OSCE, *Medical Education*, **37**, 132–139.

Livingston S.A. and Zieky M.J. (1982*). Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*: Educational Testing Service, Princeton, New Jersey.

McLaughlin, K., Gregor, L., Jones, A. and Coderre, S. (2006). Can standardized patients replace physicians as OSCE examiners?, *BMC Medical Education*, **6**:12

Newble, D. (2004). Techniques for measuring clinical competence: objective structured clinical examinations, *Medical Education*, **38**, 199–203.

Pell, G. and Roberts, T.E. (2006) Setting standards for student assessment *International Journal of Research & Method in Education*, **29(5)**, 91–103.

Regehr, G., MacRae, H., Reznick, R.K. and 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)**, 993-7.

Streiner, D.L. and Norman, G.R. (2003). *Health Measurement Scales: A Practical Guide to Their Development and Use*, 3rd edition, Oxford Medical Publications, Oxford.

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**, 115-122.

Illustrations and tables

Table 1: Assessor grading scheme for each station

Grade as shown on score sheet	Interpretation	Coded grade in data file
E	Clear fail	0
D	Borderline	1
C	Clear pass	2
B	Very good pass	3
A	Excellent pass	4

Table 1 showing the grading scheme used by assessors to award the overall station grade for a student. The final column shows how the grade is coded numerically for the analysis.

Table 2: Rating schema for simulated patients

Rating in answer to this statement: "I felt the student showed respect for me and responded to my concerns and questions in a professional manner"		
Rating awarded	**Interpretation**	**Coded rating in data file**
1	Strongly disagree	0
2	Disagree	1
3	Neutral	2
4	Agree	3
5	Strongly agree	4

Table 2 showing the rating scheme used by simulated patients to assess their opinion of the student performance at each of the relevant stations.

Table 3: Overview of the assessment data

Year	Number of students	Number of stations	Number of stations with simulated patient ratings data
2006	264	18	7
2007	267	18	10

Table 3 gives an overview of the datasets employed – the OSCE marks and grades for Year 4, 2006 and 2007.

Figure 1: The borderline method calculation of the station pass mark

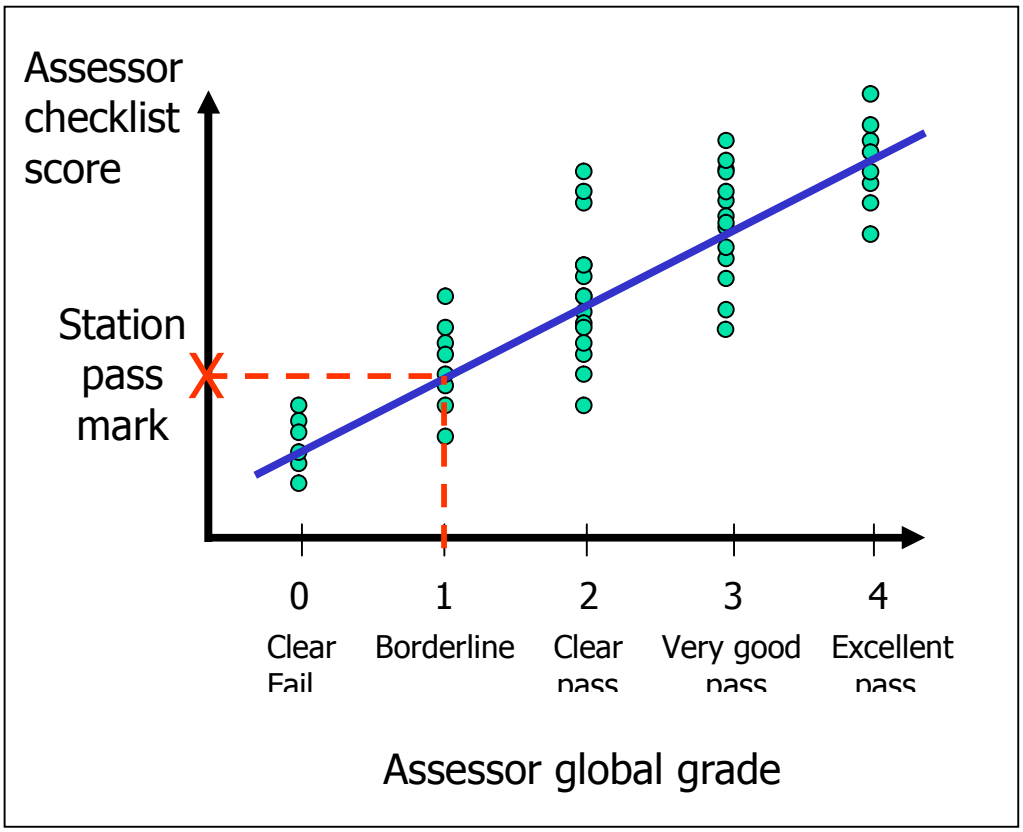

Figure 1 showing, in schematic terms, how a linear regression technique of assessor checklist scores regressed on assessor grades is used to calculate the pass mark at each individual station.