QUANTIFYING THE BORDERLINE CANDIDATE IN STANDARD SETTING

J. McLachlan[1], A. Robertson[2], B. Weller[1], M. Sawdon[3]

[1]University of Central Lancashire (UNITED KINGDOM)
[2]Cumbria, Northumberland Tyne and Wear NHS Foundation Trust (UNITED KINGDOM)
[3]University of Sunderland (UNITED KINGDOM)

Abstract

Background:

Conceptualising the Borderline candidate is one of the most difficult tasks in standard setting. However, it is also central to the process. Here we describe a methodology by which the score of Borderline candidates can be retrospectively calculated from the Facility (the percentage of items answered correctly) of assessment items for the cohort as a whole.

Methods:

We previously explored performance of candidates within an academic year in one UK medical school, covering 26 separate assessments. Each assessment had previously been standard set by either Angoff or Borderline Regression methods. In this study, we identified Borderline candidates by reviewing their performance within a particular test, not part of the previously published material. A student was classed as 'Borderline' if they were within 1 Standard Error of Measurement above or below the pass cut score. We plotted the item scores of the Borderline candidates as calculated by this method in comparison with Facility for the whole cohort and fitted a curve to the resulting distribution. In this paper, a simple method of repeating this process is described for any cohort of students.

Results:

For an ideal cohort of candidates, Borderline candidate scores should intercept the self-plot of all candidate scores at two places - at a facility of 100% and a facility of 20%. These correspond to all candidates getting the item correct and all candidates guessing the outcome. We observed a strong curvilinear distribution showed by Borderline candidates compared to the whole cohort. This relationship was well described by an exponential of the form $y \approx C \cdot \exp(F \cdot x)$, where y is the Facility of Borderline candidates on that Item, x is the observed Item Facility of the whole cohort, and C and F are constants.

In our previous study we had found C and F had similar values under different conditions. Ideal values for C and F of 12.3 and 0.021, intercept the self-plot of item Facilities very close to 100% and 20%. In this study, we again observed values for C and F close to these ideal values: C = 10.06 and F = 0.0231. Differentiating the equation indicates where the assessment ought to be most sensitive.

Differentiating the ideal curve of the difference between all candidates and Borderline candidates suggests an item facility at which the sensitivity of discrimination between the cohort and the borderline candidates is at a maximum. This value is approximately 64.5%.

Conclusions:

This approach can be used to standard-set assessments in their entirety when they are low stakes or norm referenced, in preference to Cohen methods. While Cohen methods are based on the performance of one candidate (or a very small number of candidates), this exponential method is based on all candidates and all items and is therefore more robust. In high stakes assessments, it can be used to correct values where the Facility is very different from the standard-set value, and its use in this context for the UK General Medical Council proposed national exam. It could also be used to standard set novel items such as Very Short Answer formats, where standard setting panels are unfamiliar with the expected performance of these items. And finally, it can be used to suggest the Item Facility at which the discrimination of a test is optimal, namely at 65.5%.

Keywords: Standard-setting, borderline, minimally competent candidate, Angoff

1 INTRODUCTION

Angoff standard setting methodology is widely used internationally [1]. However, one of its particular challenges is conceptualising the construct of the Borderline candidate [2], widely recognized as one of the most difficult tasks in standard setting. It is particularly challenging if the assessors are new to the task, and/or unfamiliar with the particular cohort of candidates under consideration. The need to review each item by a minimum number of assessors, generally twice, with discussion and negotiation of outcomes, is time-consuming, and hence expensive. However, it is also central to the process.

We have previously described a methodology by which the score of borderline candidates can be retrospectively calculated from the Facility (the percentage of items answered correctly) of assessment items for the cohort as a whole [3] (in a test employing Single Best Answer Multiple Choice Questions). This was done by first identifying Borderline candidates by their performance across the entire academic year, then exploring their performance on each Item in a given assessment in comparison to the whole cohort. As hypothesized, this forms a curve of exponential form, which is very similar in each assessment to which it is applied, and is of the general form $y \approx C \cdot \exp(F \cdot x)$, where y is the Facility of Borderline candidates on that Item, x is the observed Item Facility of the whole cohort, and C and F are constants. We identified values of C and F which intercept the self-plot of all items very close to 100% and 20%, as would be required of such an ideal curve, since if all candidates answer all items correctly, all candidates *and* Borderline candidates will score 100%, while if none of the candidates can answer the item correctly, all candidates *and* Borderline candidates will score 20% on average. These values are $C = 12.3$ and $F = 0.021$.

In this paper, we confirm and extend these findings in a different group of candidates and describe a simple method by which it can be confirmed by others in any assessment of reasonable size. We also explore the significance of the curve which represents the differential of the self-plot of all results and our calculated exponential curve and determine where the resulting curve reaches a minimum rate of change. We suggest that at this point, the curve is at maximum sensitivity to distinguish between borderline candidates and the cohort as a whole.

2 METHODOLOGY

The full methodology is described in our previous paper [3]. However, we appreciate that this method is time consuming, and offer in this paper a simple way in which our results can be tested for any test employing Single Best Answer Multiple Choice Questions, where the outcomes are 0 or 1. The requirements for any such test are merely that it is of a reasonable standard in terms of the number of items and the number of candidates – we suggest perhaps 100 of each. Below these values, the signal-to-noise level may be too low. Since this description is intended to confirm the basic hypothesis, the test should have already been standard set by some professional means. Our findings are derived from Angoff standard setting procedures, and it will be most instructive to learn if they also apply to situations where different standard setting methods are used.

All analyses were carried out in Excel, to ensure that they could readily be repeated without the requirement of particular statistical expertise or software.

First, the results of the exam are tabulated, with candidates represented in the first column, and the outcome for each item competed in the corresponding rows. Candidates can be represented by arbitrary code numbers to preserve confidentiality, if it is desired to publish the results of any such studies, or to compare several different tests. The last column should represent the percentage total correct for each candidate. Such tables can readily be exported from various exam software packages, for instance, Maxinity Maxexam © [4], using the ‘Export’ function.

Second, the table is sorted from highest to lowest percentage scores.

Third, the Borderline candidates are identified as all those within plus or minus 1 Standard Error of Measurement of the calculated cut score.

Fourth, for each item, the percentage correct scores of (a) all candidates (i.e. the Facility of the item) and (b) all Borderline candidates are calculated, and plotted in such a way that the all candidate scores

represent a self-plot, of these scores against themselves, therefore forming a 45° line on the graph (see Fig 1).

Finally, the exponential trendline can be added to the Borderline candidate points (See Fig 1).

Next, we calculated the ideal difference between all candidate Facilities (where y=x), and those for Borderline candidates (where y = 12.3·exp(x·0.021) and differentiated the resulting equation. Results are shown in the succeeding sections.

3 RESULTS

Fig 1 shows the outcomes of steps 1 to 4 in the Methods, with one particular test with over 100 candidates and 100 MCQ Items. The exponential trendline for the Borderline candidates is shown, with y = 10.06·exp(0.023·x), which is close to our theoretical ideal curve for Borderline candidates. The R^2 value is respectable under the circumstances.

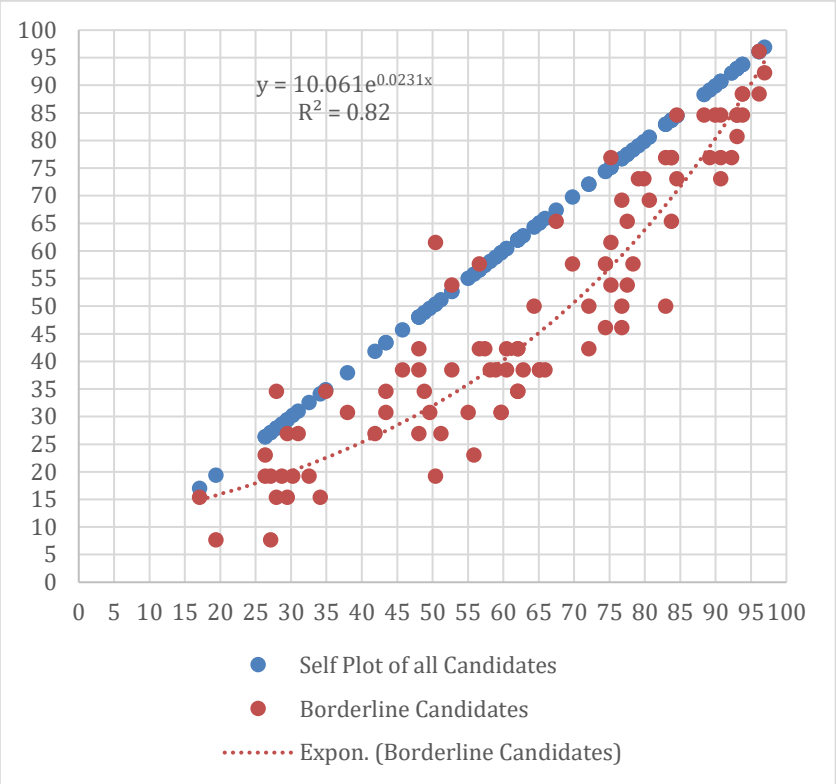

Figure 1. Item Facilities for all Candidates and Borderline Candidates

Next, we calculated the difference between the self-plot of all candidates and the ideal curve for all Borderline candidates. The equation for this is $y = x - 12.3.e^{0.021x}$

$\therefore dy/dx = 1 - (12.3*0.021)* e^{0.021x} = 1 - 0.2583* e^{0.021x}$

The maximum is when the derivative is zero, hence at that point

$1 = 0.2583\ e^{0.021x}$

$\therefore e^{0.021x} = 3.8715$

$\therefore 0.021*x = \ln(3.8715) = 1.356$

$\therefore x = 1.3536/0.021 = 64.46$

The curve is shown in Fig 2.

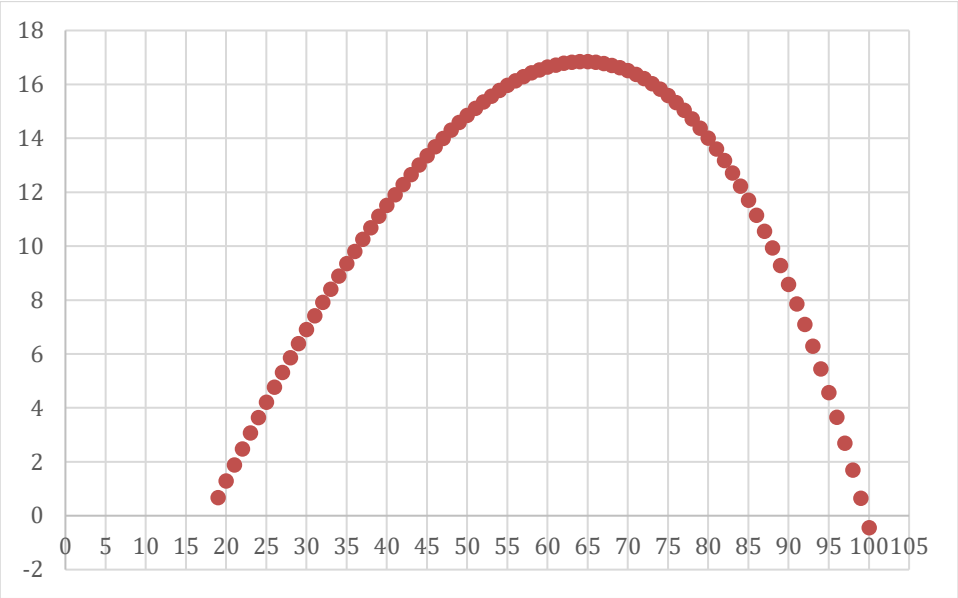

Figure 2. Plot of the difference between the self-plot of all candidates and the ideal curve for all Borderline candidates.

Since this difference is at a maximum at 64.46% (rounding to 64.5% for practical purposes) this suggests that where Item Facility is at this value, then the discrimination between all candidates and Borderline candidates is greatest.

4 CONCLUSIONS

This approach could be used to standard-set assessments in their entirety when they are low stakes, in preference to Cohen methods. While Cohen methods are based on the performance of one candidate (or a very small number of candidates), this exponential method is based on all candidates and all items and is therefore likely to be more robust. It should be noted, however, that in the end it would, like the Cohen methods, reduce to a norm-referenced approach, and therefore should be used with caution for high stakes assessment. Having said that, however, a norm refenced approach may be appropriate In high stakes progress testing methodologies [5].

However, the method could also be in high-stakes testing to correct values for individual items where the Facility is very different from the standard-set value, and its use is planned in this context for the UK General Medical Council's proposed national Medical Licensing Assessment [6], where it will be compared with Item Response Theory methods [7]. At the moment, items which perform very differently from their predicted Angoff value, may be removed from the assessment, and candidates may therefore lose or gain a score point, depending on which approach is taken [8] (and neither of which is particularly satisfactory).

It could also be used to standard set novel items such as Very Short Answer formats [9], where standard setting panels are unfamiliar with the expected performance of these items. These have lower Facility than otherwise very similar SBAs from which they have been derived, showing that Angoff values would be inappropriate for use in this context.

Finally, an implication of the form of the ideal curve, is that it suggests an Item Facility at which the discrimination of the item would be optimal, namely at 65.5%. In principle, a test could be designed where all the items were around this value, where item performance statistics from previous administrations were available. Alternatively, this value could be used as the start point in Computer Adaptive Testing [10].

ACKNOWLEDGEMENTS

Grateful thanks to Chris McManus, Paul Tiffin and anonymous contributors on a scientific forum for valuable discussions.

REFERENCES

References [Arial, 10-point, left alignment, upper and lower case] should be cited according to the Bibliography and Citation Style https://iated.org/citation_guide

[1] Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL, editor. Educational measurement. 2nd ed. Washington, DC: American Council on Education; 1971. p. 508–600 A. Einstein, “General theory of relativity,” *Annalen der Physik*, vol. 49, no. 7, pp. 769–822, 1916.

[2] Clauser JC, Hambleton RK, Baldwin P. The effect of rating unfamiliar items on Angoff passing scores. Educ Psychol Meas. 2017;77:901–16. A.A. Author, "Journal/Conference Article Title," *Periodical Title*, vol. Volume, no. Issue, pp.-pp., Publication Year.

[3] McLachlan JC, Robertson KA, Weller B, Sawdon M. (2021) An inexpensive retrospective standard setting method based on item facilities. *BMC Medical Education* 21:7 https://doi.org/10.1186/s12909-020-02418-5

[4] Maxinity Mexexam.

[5] McHarg J, Bradley P, Searle J, Ricketts C, Chamberlain S, McLachlan JC (2005) Assessment of progress tests. *Medical Education* **39**:221-227

[6] MLA

[7] Mark Gurnell, personal communication

[8] Tavakol, M. and Doody, G.A., 2015. Making students' marks fair: standard setting, assessment items and post hoc item analysis. *International journal of medical education*, *6*, p.38.

[9] Sam, A.H., Field, S.M., Collares, C.F., van der Vleuten, C.P., Wass, V.J., Melville, C., Harris, J. and Meeran, K., 2018. Very-short-answer questions: reliability, discrimination and acceptability. *Medical education*, *52*(4), pp.447-455.

[10] Gershon, R.C., 2005. Computer adaptive testing. *Journal of applied measurement*.