Food for Thought: A New Compromised Method of Standard Setting

Reza Vahid Roudsari[1]

Corresponding author: Mr Reza Vahid Roudsari reza@manchester.ac.uk
Institution: 1. The University of Manchester
Categories: Assessment, Selection

Received: 20/09/2016
Published: 23/09/2016

Abstract

A number of standard setting methods are in use with their own strengths and limitations. Angoff is a popular method worldwide but very demanding on resources. This paper explores if the judgement of the Angoff panels of judges can been explained by post-hoc psychometrics. Furthermore the paper examines if such correlations can be used to standard set exam papers based on their psychometric item difficulty and discrimination index.

Keywords: standard setting, progress testing, psychometrics

Introduction

Methods of standard setting have evolved during the past five decades, ensuring examinees experience fairer examinations. The literature on the validity and reliability (Bandaranayake, 2008; Ben-David, 2009; McKinley & Norcini, 2014; Norcini, 2003; Verhoeven et al., 1999) of these methods is enriching on annual basis, making these methods more defensible (Boulet, De Champlain, & McKinley, 2003; Downing, Tekian, & Yudkowsky, 2006), acceptable (Cizek, 1996) and economical (Norcini, 2003). Looking at the available evidence, such methods can be broadly divided into three categories of absolute, relative and compromised.

Absolute methods of standard setting have gained increased popularity in the fields of medical and dental education: they are easy to understand and adopt, do not require complex mathematical calculations and are defensible if challenged. The Angoff method (Angoff, 1971), for example, works on the basis of the prediction of the performance of a pre-defined group of borderline passing candidates (BPCs) for a given examination paper. In this method, the characteristics of the BPCs are explained to the panel of the judges. The judges would discuss individual question items and submit their prediction of performance for that given cohort of BPCs.

The Ebel method (Ebel, 1971, 1972), however, works very similar to the Angoff method in the setup but with the major difference that the judges will not only look at the ‘difficulty’ of individual test items but also the ‘importance’
of each item for the given cohort of students. A pre-defined matrix is then used to identify the pass mark based on the data from the judges.

Item difficulty, however, is a test psychometrics known to the field of education for over half a century (Goodwin, 1999). There have been several attempts in the past to demonstrate a correlation between judges’ perception of the item difficulty with the post-hoc psychometric item difficulty (Goodwin, 1999; Impara & Plake, 1998); but the evidence so far is not convincing and, in the most optimistic way, moderate correlation could be shown.

Having said that, one can question our standard setting methods: if the judges cannot predict the difficulty of an item are they setting the pass mark correctly? To add to the drawbacks of these methods one can argue that these methods are time consuming, requiring considerable time input and commitment from mostly senior academic members of staff. Literature recommends 10-15 judges per panel of examination as an optimal number (Fowell, Fewtrell, & McLaughlin, 2008; Hurtz & Hertz, 1999), which for most medical and dental schools this is a large stretch on their already scarce resources.

The relative methods of standard setting have undergone major development since the introduction of the Contrasting Groups (Huynh, 1976) and Borderline Regression (Schoonheim-Klein et al., 2009) methods; however their use is somehow limited to observed methods of assessment and cannot be utilized in the common written formats of assessment, Multiple Choice Question papers for example. Other relative methods of standard setting, for example Cohen (Cohen-Schotanus & van der Vleuten, 2010) and Wijnen (van der Vleuten, Verwijnen, & Wijnen, 1996) have great place of use in admission examinations but have major drawbacks when used for progression tests.

The compromised methods of standard setting have tried to incorporate a number of good properties of the above two major categories while attempting to overcome some of their drawbacks. The Hofstee method (Hofstee, 1983), for instance, is very economical to run, requires minimal input from the judges and is defensible; however, the judges commit themselves to suggesting ‘minimal’ and ‘maximum’ failure rates that may not sound very fair to the examinees, especially on a progression exam setting.

As mentioned earlier, there have been some attempts to explore the relationship between the psychometric difficulty of the test items with the standard setting prediction of the judges. It is, however, useful to be reminded that the item difficulty is one of many possible post-hoc psychometric tests. There is a possibility that such correlation may still exist but not in a single dimensional way as explored by previous researchers. For this reason, the author has explored the possibility of such a correlation when more than one variable is introduced into the equation within the limits of the readily available data to him.

The aim of this paper is to explore the possibility of setting a pass score using the pattern of the ‘judgment’ of the judges and a number of post-hoc psychometric tests.

The setting

The School of Dentistry (SoD) of the University of Manchester runs a five-year BDS programme. The first two years of the programme consist of a split of modules run by the Faculty of Life Sciences (FLS) as well as the SoD. The SoD runs the final three BDS years. The School has adopted several methods of assessment and uses a number of different standard setting methods appropriate to the assessment tools used.

The Single Best Answer (SBA) paper is one of the assessment methods that is run across all of the five BDS years;
however, a large number of questions in the first two BDS years are provided and standard set by the FLS. Since two separate panels of standard setting are used in the first two BDS classes, their data was not considered for the purpose of this paper.

A total of 245 students, enrolled on the years 3, 4 and 5 of the BDS programme, sat their SBA exam papers in May 2016. Each BDS paper consisted of 120 question items. Each question item consisted of a vignette and five options, one of which was the single best answer. The question items were selected by their BDS year leads based on the examination blueprint from the SBA question banks. The blueprint for each BDS year covers a wide range of domains according to the intended learning outcomes set by the General Dental Council (GDC, 2011).

The SBA exams are held electronically at the University computer cluster facilities using the virtual learning environment (Blackboard Inc. Released 2013. Blackboard Version 9. London, UK) and are invigilated by the University members of staff. All the SBA exams for the three BDS classes were run in the second week of May 2016. Parallel to this process, three separate panels proceeded with the standard setting of each exam paper using the Angoff method. Each panel consisted of four experienced members of staff and two part-time clinical tutors, all involved in the teaching of the students on the clinic.

The School Assessment Lead (RVR) collected all the exam data as well as the data from the standard setting panels for analysis and processing. Each exam paper was inspected using the Discrimination Index (DI) and Item Difficulty (Diff) post-hoc. Between the three papers, a total of 5 questions had to be remarked due to the wrong key being entered and 1 question had to be omitted as a critical piece of information was missing in the vignette. Using the classic test theory, the reliability measure for each exam paper (Chronbach’s alpha or Chα) as well as the Standard Deviation (SD) and Standard Error of Measurement (SEM) were calculated (Table 1).

| Cohort (May 2016) | Angoff (%) | SD  | Chα  | SEM  |
|------------------|------------|-----|------|------|
| BDS Y3           | 49.66      | 10.210 | 0.742 | 5.19 |
| BDS Y4           | 56.04      | 6.986  | 0.550 | 4.69 |
| BDS Y5           | 62.90      | 7.181  | 0.760 | 3.52 |

Regression analysis

SPSS (IBM Corp. Released 2013. IBM SPSS Statistics for Macintosh, Version 22. Armonk, NY) was used for all the data processing. For each exam paper, Diff by DI scattered plots were drawn and cubic (cubic) and quadric (quad) regression curves were calculated and illustrated. The Angoff value for each BDS paper was marked on the Y-axis and their intersects with the cubic (x_{cfit}) and quad (x_{qfit}) regression curves were calculated on the X-axis (Figure 1 and Table 2).
Figure 1. The cubic (A) and quad (B) regression curves drawn for the BDS Y3. The Angoff pass mark (49.66%) is marked on the Y-axis and its intersect with the X-axis calculated.
Table 2. The intersect of the Angoff standard set on the cubic and quad regression curves are calculated on the X-axis, \( x_{cfit} \) and \( x_{qfit} \) respectively.

| Cohort   | \( x_{cfit} \) | \( x_{qfit} \) |
|----------|----------------|----------------|
| BDS Y3   | -0.0721        | -0.0585        |
| BDS Y4   | -0.1111        | -0.1106        |
| BDS Y5   | -0.0674        | -0.0639        |

Correlation between the variables

The inter-item correlation between the \( x_{cfit} \) and \( x_{qfit} \) with the SD, Ch\( \alpha \) and SEM were explored (Table 3). From the three psychometric variables, Ch\( \alpha \) had the highest inter-item correlation to the \( x_{cfit} \) and \( x_{qfit} \). The author refrained from running any statistical tests due to small sample size.

Scattered plots were drawn to illustrate the correlation between the Ch\( \alpha \) to \( x_{cfit} \) and \( x_{qfit} \). Best fit correlation lines were drawn and their linear regression equations calculated (Figure 2). Based on this data, the \( x_{cfit} \) and \( x_{qfit} \) can be calculated using the following equations:
Equation (i): \( x_{cfit} = -0.22 + 0.21 \times Ch\alpha \) and

Equation (ii): \( x_{qfit} = -0.24 + 0.24 \times Ch\alpha \)

Table 3. Correlation between the \( x_{cfit} \) and \( x_{qfit} \) variables with SD, Ch\( \alpha \) and SEM of the exams.

|        | SD    | Ch\( \alpha \) | SEM   |
|--------|-------|----------------|-------|
| \( x_{cfit} \) | 0.461 | 1.000          | -0.320 |
| \( x_{qfit} \) | 0.623 | 0.985          | -0.133 |

Figure 2. The correlation of cubic (A) and quad (B) xfit points (\( x_{cfit} \) and \( x_{qfit} \) respectively) to Ch\( \alpha \) was assessed using linear regression.

A

B
Hypothesis

The author hypothesizes that the Angoff standard setting could have been predicted using the cubic and quad regression curves drawn based on the Diff of the question items by their DI. The vertical line that intersects with these curves has a value on the X-axis that can be calculated with equations (i) and (ii) for the cubic and quad methods respectively. The later value is correlated to the reliability of the exam, herein we used the Chronbach's alpha.

For example, the standard set pass mark for BDS year 3 could have been calculated with the cubic regression method ($S_{cubic}$) using the following equation (also note Figure 1):

**Equation (iii):**  
$$S_{cubic} = 0.58 + 0.94x_c - 2.8x_c^2 + 2.62x_c^3$$

where $x_c$ is calculated using **equation (i):**  
$$x_c = -0.22 + 0.21 \cdot Ch\alpha$$

The author suggests that the equations (i) and (ii) reflect the 'judgment' of the standard setting panels in our School and should remain valid for all of the diets of our exams.
Analysis of the hypothesis

Based on the above hypothesis, the new standard set values using the cubic ($St_{cubic}$) and quad ($St_{quad}$) regression methods were calculated for the three BDS classes who sat the May 2016 exams (Table 4) and illustrated in Figure 3. It was not appropriate to run any statistical tests due to the small number of the samples.

To further explore the performance of the BDS classes based on the new methods of standard setting, these values were applied to investigate how they influence the overall failure rate of each given class (Table 5). Visual inspection of this data suggests that the cubic regression method would have resulted in very similar outcome to the Angoff method.

Table 4. $St_{cubic}$ and $St_{quad}$ calculated for each of the BDS years for the May 2016 diet.

| Cohort (May 2016) | Angoff | Chα | $x_c$ | $St_{cubic}$ | $x_q$ | $St_{quad}$ |
|------------------|--------|-----|-------|--------------|------|------------|
| BDS Y3           | 49.66  | 0.742 | -0.06418 | 50.74 | -0.06192 | 49.33      |
| BDS Y4           | 56.04  | 0.550 | -0.10450 | 57.44 | -0.10800 | 56.44      |
| BDS Y5           | 62.90  | 0.760 | -0.06040 | 64.02 | -0.05760 | 61.46      |

Figure 3. Graphic illustration of the standard set values (%) using the Angoff, cubic and quad regression methods for the May 2016 sitting of the exams.
Table 5. The failure number (and rate) for each BDS class according to the different methods of standard setting used in the May 2016 sitting.

| Cohort (May 2016) | Number of students | Angoff (rate %) | Cubic (rate %) | Quad (rate %) |
|-------------------|--------------------|-----------------|----------------|--------------|
| BDS Y3            | 76                 | 5 (7%)          | 5 (7%)         | 4 (5%)       |
| BDS Y4            | 83                 | 0 (0%)          | 0 (0%)         | 0 (0%)       |
| BDS Y5            | 86                 | 5 (6%)          | 6 (7%)         | 1 (1%)       |

Exploring the validity of the new methods

To explore if the new methods could have predicted the pass mark of the previous diets of the exams, the author analyzed the raw exam data available for the May 2015 sitting of the exams. If equations (i) and (ii) truly reflect the ‘judgment’ of our panels, they should remain true when they are used to re-calculate the standard set of the previous
year exams.

For this reason, Diff and DI for the question items of the previous exam papers were calculated. The cubic and quad regression curves were drawn and their equations calculated for each BDS year. The Chα value for each exam paper was calculated. The St_{cubic} and St_{quad} were calculated using equations (i) and (ii) with their respective Chα. The stored Angoff data for May 2015 diet were acquired. This data is presented in Table 6. The standard set values using the three methods (Figure 4) and the failure rates using each method (Table 7) are presented.

Once again there is close resemblance of the outcomes from cubic regression method to the Angoff. The author suggests that it might be possible to use the cubic regression method to safely substitute the Angoff method of standard setting; however, due to the small sample size, statistical testing in this paper was not appropriate. Further research by institutions with a large number of samples readily available to them is required to approve or reject this hypothesis.

Table 6. St_{cubic} and St_{quad} calculated for each of the BDS years for the May 2015 diet.

| Cohort (May 2015) | Angoff | Chα | x_c | St_{cubic} | x_q | St_{quad} |
|------------------|--------|-----|-----|------------|-----|-----------|
| BDS Y3           | 52.96  | 0.728 | -0.06712 | 63.38 | -0.06528 | 63.56 |
| BDS Y4           | 52.44  | 0.699 | -0.07321 | 51.86 | -0.07224 | 63.04 |
| BDS Y5           | 52.85  | 0.654 | -0.08266 | 57.23 | -0.08304 | 54.79 |

Figure 4. Graphic illustration of the standard set values (%) using the Angoff, cubic and quad regression methods for the May 2015 sitting of the exams.
Table 7. The failure number (and rate) for each BDS class according to the different methods of standard setting used in the May 2015 sitting.

| Cohort (May 2015) | Number of students | Angoff (rate %) | Cubic (rate %) | Quad (rate %) |
|-------------------|--------------------|-----------------|----------------|--------------|
| BDS Y3            | 84                 | 0 (0%)          | 7 (8%)         | 7 (8%)       |
| BDS Y4            | 86                 | 0 (0%)          | 0 (0%)         | 5 (6%)       |
| BDS Y5            | 67                 | 0 (0%)          | 0 (0%)         | 0 (0%)       |

Discussion

Angoff and Ebel are two popular methods used to standard set multiple choice question items. They have gained worldwide popularity, not only because they are easy to understand and adopt, but also they require minimal training and calibration of the judges as well as minimal mathematical knowledge to produce results. On the other hand, however, they drain valuable academic resources if done properly. Finding a mutual time to gather over 10 academic members of staff, mostly in senior positions, is not an easy task. If the process is done remotely, as it is done in our SoD, there will be question items that require discussion and the time saved by the remote process is spent on
discussing the items with high variability between the judges’ scores.

Understandably, these methods become hard to defend when the number of the judges on the panel is reduced. The ‘Hawk and Dove effect’ can potentially skew the pass mark to one way or another, either way making the outcome unfair to the examinees and potentially making the examination process susceptible to legal challenges. The compromised methods of standard setting, on the other hand, aim to take advantage of the ‘relativeness’ of the data produced by the performance of the cohorts of the examinees and at the same time aim to save the time input required by the expensive and busy academic members of staff. In this paper the author hypothesizes that the judgment of our panels of standard setting can be explained by a linear regression that happened to be correlated to the test of the reliability of our exams. This data is produced based on the behavior of our judges as well as the psychometrics of our exam data in the May 2016 diet.

Previous authors had unsuccessful attempts to find a relationship between the psychometric difficulty of the question items and the standard setting prediction of the judges; however, it can be argued that this is not just the difficulty of a question item that matters. The judges are unintentionally influenced by the discrimination factor of each item; i.e. who finds the question item difficult. For this reason the author explored the data not in a single dimensional way but instead investigated the interaction between the psychometric item difficulty value and its discrimination index. In this paper such relationships were best illustrated by cubic or quadric curves; although their correlation coefficients were low. The author then used these curves to calculate a pass mark very close to those suggested using the Angoff method. For doing so, the author suggested that the judgment of the panels of the judges could be explained using a linear regression model that is correlated to the Chronbach’s alpha of the exam paper.

The author, however, is aware that to prove or reject a hypothesis, appropriate statistical tests are essential. Based on the data gathered from the May 2016 diet of the exams, data for 25 exam papers are required to statistically distinguish 1.5% difference in the standard setting pass mark between the cubic method versus the Angoff method with 95% confidence at 80% power. Although the SoD retains the processed exam data for decades; however, the raw exam data essential to run the required statistical and psychometric tests were only available to the author from the May 2015 diet. It will take the author just over 7 years to accumulate enough exam data before any meaningful statistical tests can be run; therefore the author has decided to share this hypothesis with the community of medical and dental educators who may already have enough samples to run such tests or have access to frequent exam diets that can produce such data in a short span of time.

The author hypothesizes that the judgment of the standard setting panels in our SoD can be explained by the equations (i) and (ii). Readers have to be reminded that these equations work based on the training and calibration provided in the author’s institution. This will be different to other schools with different training, calibration and of course different cohort of students on different curriculums. The readers will need to calculate their own regression lines based on at least one year’s worth of exam data. Once these equations are calculated, they can be used to calculate the pass mark of prospective or retrospective exam diets and subsequently compared to other methods of standard setting in use.

The author used the data from three exam papers all run in May 2016. In author’s experience, the equations could be best described using a linear model; however, such relationship may still exist but explained by a curve in other institutions. Furthermore, the three data points that were used to form this line belong to three different classes of BDS. It is possible that stronger relationships could be found if such regression lines were calculated for a single class of BDS over time; the data from the BDS year 5 only, for example, gathered over the past three diets. Unfortunately the author does not have access to enough raw data to test this but invites the community to explore such relationships.
Finally, the validity of methodology of this paper can be questioned as the Angoff method of standard setting was used as the gold standard and all the comparisons are made against it. One may argue that with only six judges in each panel, our standard set pass mark may have been deviated from the true pass score for the paper. It has been shown that the calibration and training of the judges can result in reduction of the number of the judges (Fowell et al., 2008) but the above argument still remains a possibility. When comparing the outcomes of the different methods, the May 2015 diet of exams for the BDS year 3 class showed the most significant discrepancy. **Figure 5** illustrates the performance of this cohort as well as the cut off points for two of the standard setting methods. Did the performance of this given cohort exceed our expectations or did our judges behave too generously? This is a question that remains impossible to answer.

**Figure 5.** Graphic illustration of the performance of the BDS year 3 class in May 2015 diet of the exam. The Angoff (solid) and cubic (dots) pass marks are shown.

The author hypothesizes that the Angoff standard setting outcome may be predicted using the cubic and quad regression curves drawn based on the Item Difficulty of the Single Best Answer question items by their Discrimination Index values. The vertical line that intersects with these curves has a value on the X-axis that can be

---

**Take Home Messages**

The author hypothesizes that the Angoff standard setting outcome may be predicted using the cubic and quad regression curves drawn based on the Item Difficulty of the Single Best Answer question items by their Discrimination Index values. The vertical line that intersects with these curves has a value on the X-axis that can be
calculated using mathematical methods. This is due to the high correlation between these values and the test of the reliability of the exam, herein we used the Chronbach's alpha. The author is inviting the community of medical and dental educators who have access to sufficient exam data to test the validity of this hypothesis.

Notes On Contributors

The author is a clinical lecturer and the School Lead for Assessment on the BDS and BSc programmes.

Acknowledgements

Bibliography/References

Angoff, W. H. (1971). Scales, Norms, and Equivalent Scores. In T. R. L (Ed.), Educational Measurement (pp. 514-515). Washington, DC: American Council on Education.

Bandaranayake, R. C. (2008). Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. Medical Teacher, 30(9-10), 836-845.

http://dx.doi.org/10.1080/01421590802402247

Ben-David, M. F. (2009). AMEE Guide No. 18: Standard setting in student assessment. Medical Teacher, 22(2), 120-130.

http://dx.doi.org/10.1080/01421590078526

Boulet, J. R., De Champlain, A. F., & McKinley, D. W. (2003). Setting defensible performance standards on OSCEs and standardized patient examinations. Medical Teacher, 25(3), 245-249.

http://dx.doi.org/10.1080/0142159031000100274

Cizek, G. J. (1996). Standard-Setting Guidelines. Educational Measurement: Issues and Practice, 15(1), 13-21.

http://dx.doi.org/10.1111/j.1745-3992.1996.tb00802.x

Cohen-Schotanus, J., & van der Vleuten, C. P. (2010). A standard setting method with the best performing students as point of reference: practical and affordable. Medical Teacher, 32(2), 154-160.

http://dx.doi.org/10.3109/01421590903196979

Downing, S. M., Tekian, A., & Yudkowsky, R. (2006). Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. Teaching and Learning in Medicine, 18(1), 50-57.

http://dx.doi.org/10.1207/s15328015tlm1801_11

Ebel, R. L. (1971). Criterion-referenced measurements: Limitations. The School Review, 79(2), 282-288.
Ebel, R. L. (1972). Essentials of Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall.

Fowell, S. L., Fewtrell, R., & McLaughlin, P. J. (2008). Estimating the minimum number of judges required for test-centred standard setting on written assessments. Do discussion and iteration have an influence? Advances in Health Sciences, 13, 11-24.

GDC. (2011). Preparing for practice. Retrieved from http://www.gdc-uk.org/Newsandpublications/Publications/Publications/GDC Learning Outcomes.pdf

Goodwin, L. D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. Applied Measurement in Education, 12(1), 13-28.

Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson & J. S. Helminck (Eds.), On Educational Testing (pp. 109-127). San Francisco: Jossey-Bass.

Hurtz, G. M., & Hertz, N. R. (1999). How Many Raters Should be Used for Establishing Cut off Scores with the Angoff Method? a Generalizability Theory Study. Educational and Psychological Measurement, 59(6), 885-897.

Huynh, H. (1976). Statistical consideration of mastery scores. Psychometrika, 41(1), 65-78.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. Journal of Educational Measurement, 35(1), 69-81.

McKinley, D. W., & Norcini, J. J. (2014). How to set standards on performance-based examinations: AMEE Guide No. 85. Medical Teacher, 36(2), 97-110.

Norcini, J. J. (2003). Setting standards on educational tests. Medical Education, 37(5), 464-469.

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.
Van der Vleuten, C. P. M., Verwijnen, G. M., & Wijnen, W. H. F. (1996). Fifteen years of experience with progress testing in a problem-based learning curriculum. Medical Teacher, 18(2), 103-109.

http://dx.doi.org/10.3109/01421599609034142

Verhoeven, B. H., Van der Steeg, A. F. W., Scherpbier, A. J. J. A., Muijtjens, A. M. M., Verwijnen, G. M., & van der Vleuten, C. P. M. (1999). Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. Medical Education, 33(11), 832-837.

http://dx.doi.org/10.1046/j.1365-2923.1999.00487.x

Appendices

Declaration of Interest

The author has declared that there are no conflicts of interest.