International Journal of Learning, Teaching and Educational Research
Vol. 4, No.1, pp. 1-26, April 2014

Angovian Methods for Standard Setting in Medical Education: Can They Ever Be Criterion Referenced?

Brian Chapman
School of Rural Health (Churchill)[1]
Faculty of Medicine, Nursing and Health Sciences
Monash University
Churchill, Victoria, Australia

Abstract. This paper presents a discussion of Angovian methods of standard setting – methods which are widely used with the intent of defining criterion-referenced absolute standards for tests in medical education. Most practitioners, although purporting to pursue absolute, criterion-referenced standards, have unwittingly slipped into focussing on norm-referenced concepts of 'borderline' students and their predicted ability to answer assessment items in a test. This slippage has been facilitated by a shift in language from the original concept of 'minimally acceptable' persons to the modern concept of 'borderline' persons. The inability of university academics to predict accurately the performance of 'borderline' graduate-entry medical students is illustrated by presentation of data obtained from three successive cohorts of a small regional medical school during the years 2010-2012. Other data are presented to show how student performance, both 'borderline' and general, can be significantly altered by switching from didactic lectures to tutorials preceded by task-based active learning.

A protocol, based on a stricter interpretation of what is meant by a 'minimally acceptable' person, is suggested for moving towards a more criterion-referenced standard for a test based on the curriculum's learning objectives. Nonetheless, the fallibility of criterion-referenced standard-setting processes means that norm-referenced relative standards may need to be brought into play to deal with anomalous grade results should they arise. The ideal of defining an absolute criterion-referenced standard for a test, using the most commonly implemented Angovian method, is probably as least as unattainable for graduate-entry medicine as it has been previously shown to be for secondary school science.

Keywords: standard setting; Angoff method; medical education; norm-referenced standard; criterion-referenced standard

[1] Formerly Gippsland Medical School (2007-2013).

1. Introduction

Medical schools are required to define standards of quality assurance in the assessment of medical trainees such that society can have confidence in the professional competence of medical graduates once they are registered to practice. To this end, the quality of a medical curriculum is defined by the clarity and comprehensiveness of its stated learning objectives, and the efficacy of the teaching and learning processes directed towards the attainment of those objectives is assessed using a variety of measuring instruments. These instruments may include written examinations comprising multiple-choice questions (MCQs), extended matching questions (EMQs) and short-answer questions (SAQs), *viva voce* examinations such as the Objective Structured Clinical Examination (OSCE), or a variety of essays and assignments. Quality assurance then focuses on defining a minimally acceptable standard of competence for each assessment question or task encountered by the students as they progress through the course. In common with most licensing and certifying operations, it is desired to define an *absolute* standard of competence for assessing the quality of a medical graduate rather than a *relative* standard expressed by comparison either with other candidates in a given cohort or with the performance of preceding cohorts. The definition of an absolute standard is called criterion referencing while the use of a relative standard is called norm referencing; application of criterion referencing is intended to establish minimum standards of competence and this is widely held to be preferable to norm referencing (Searle, 2000; Norcini, 2003; Downing, Tekian, & Yudkowsky, 2006).

A cursory glance at the literature on assessment in medical education will reveal the widespread use of methods for standard setting attributed to William H. Angoff (1919-1993), a researcher at the Educational Testing Service in the United States for 43 years, whose main contributions to educational research and practice were focussed on the measurements used in testing and scoring. The key reference cited for this attribution is Chapter 15 'Scales, Norms, and Equivalent Scores' in *Educational Measurement, Second Edition*, edited by Robert L. Thorndike (Angoff, 1971). Yet, within this 93-page chapter, as many people have noted over the years (e.g., Zieky, 1995, pp.8-9; Cizek & Bunch, 2007, p.81), Angoff's original description is very short, comprising no more than nine sentences of text distributed between two paragraphs and an associated footnote.

The purpose of this paper is to present a critical discussion of the rationale, intent and implementation of the most widely used of the several variants of Angoff's (1971) original suggestion that have emerged. Preparation for this discussion reveals that very little can be said today in criticism of Angovian procedures that has not been said before. This suggests that much current practice is based on pragmatism, allowing standard setting to proceed for any number of reasons, including ignorance or wilful disregard of the many objections and concerns that have been raised in the past. It is not the aim of this discussion to review this literature or to rehearse old arguments. Rather, the original contribution sought here is to illuminate the discussion by identifying the problems of linguistic imprecision and conceptual vagueness that have interacted and confounded the most well-intentioned quests for defining absolute assessment standards in medical education. Samples of original assessment data are included to illuminate the discussion further.

The detailed analysis and discussion will be usefully facilitated by distinguishing two methods of standard-setting contained in Angoff's (1971) original description: Angoff's Text Method (ATM), and Angoff's Footnote Method (AFM).

2. Angoff's Text Method (ATM)

The context for Angoff's original description is the standard-setting problem of finding a suitable 'pass mark' and 'honours mark' on a scale applied to a test, such cut scores being "decided on the basis of careful review and scrutiny of the items themselves" (Angoff, 1971, p.514). The aim was to establish standards that would be independent of normative data relating to actual "performance as it exists" (p.514). This raises immediately the problem of determining whether a test is criterion referenced or norm referenced. Although Angoff doesn't express the issue in these words, it seems that he is striving to define a criterion-referenced standard that will stand immutable in the face of actual performance data. To this end, Angoff's (1971, pp.514-515) two paragraphs specify ATM as follows:

> A systematic procedure for deciding on the minimum raw scores for passing and honors might be developed as follows: keeping the hypothetical "minimally acceptable person" in mind, one could go through the test item by item and decide whether such a person could answer correctly each item under consideration. If a score of one is given for each item [p.515] answered correctly by the hypothetical person and a score of zero is given for each item answered incorrectly by that person, the sum of the item scores will equal the raw score carried by the "minimally acceptable person." A similar procedure could be followed for the hypothetical "lowest honors person."
>
> With a number of judges independently making these judgments it would be possible to decide by consensus on the nature of the scaled score conversion *without actually administering the test*. *If desired*, the results of this consensus could later be compared with the number and percentage of examinees who actually earned passing and honors grades [emphasis added].

Looking back over the ensuing four decades or more since the above words were written, it may be safely concluded that the thinking of Angoff in this specific instance, and of all those who have subsequently used Angovian methods of standard setting, has been dominated by the view that assessment "should be objective, measurable and 'certain' (and therefore that assessment can be made reliable and valid)" (Williams, 2008, p.402). This is implicit in the notion that an absolute standard for a test can be set "without actually administering the test" and that such a standard *might* be compared with "performance as it exists" … "if desired".

However, Angoff's (1971) prescription is far from being a robust example of criterion referencing in action. This is because of the different constructions that might reasonably be placed upon the first paragraph.

2.1 Angoff's First Paragraph

The wording of this paragraph, as quoted above, is insufficiently precise to yield a single, unambiguous reading. As Zieky (1995, p.9 footnote 4) has noted, Angoff's (1971) use of the word 'could' is frustrating in its lack of precision and its uncertain distinction from alternatives such as 'should' or 'would'.[2] On this matter Impara and Plake (1997, p.363 note 1) also observe that 'should' is typically interpreted as a higher target than 'would'. In the same footnote 4 of Zieky (1995, p.9), we find that, when Zieky personally asked Angoff in the early 1980s which of 'could' or 'would' was correct, Angoff "replied that he did not think it mattered very much". This is very illuminating and it suggests that Angoff did not think it important to develop a full appreciation of the importance of linguistic precision in defining an unambiguous method for establishing a criterion-referenced standard.[3] In that sense, therefore, Angoff (1971) did not set his method on a sufficiently firm foundation.

However, let us choose the least vague of the three alternatives – *would* – and see where that leads us. The criterion-referenced standard-setting prescription of ATM then becomes:

> Keeping the hypothetical "minimally acceptable person" in mind, one could go through the test item by item and decide whether such a person *would* answer correctly each item under consideration.

There remains a problem with the dual focus of the procedure. Does the focus lie on the concept of the *minimally acceptable person* or on the *content of each item*? The crucial linguistic watershed here is Angoff's (1971) concept of the "minimally acceptable person". It is possible to construct this concept in two ways, one lending itself more readily to criterion referencing than the other.

A. Firstly, it may be given a more criterion-referenced construction by implicit reverse engineering of the text. In short, *by definition*, a "minimally acceptable person" *will* answer correctly *every* item that the assessors have identified as embracing the 'minimally acceptable performance' criteria for the test. So the pass mark becomes the sum of all the marks deriving from such 'minimally acceptable performance' items. The *procedure* under this construction would be to ask of the item, not whether a "minimally acceptable person" could answer it correctly, but whether it encapsulates an element of 'minimally acceptable

[2] Angoff (1971) uses 'would' in the "slight variation" of ATM represented by AFM.

[3] As suggested later (see Footnote 8), there is no reason why Angoff (1971) *should* have given these matters any more thought or space than he did within the context of his original article. This is the responsibility of contemporary users of Angovian methods.

performance'. It seems that this construction has been attempted rarely, if at all.

B. Alternatively, it may be given a less criterion-referenced construction by allowing the *difficulty* of the item (as distinct from its criterion-referenced *content*) to weigh in the assessors' estimates as to whether or not a "minimally acceptable person" *would* answer the item correctly. This question cannot be answered with any certainty because the focus has moved away from the item's content to the ability of a "minimally acceptable person" to answer the item successfully. Any estimate of this ability must necessarily take into account a margin for error that such guesswork entails. This construction inevitably tends towards norm referencing, where the 'norm' is a person or group of persons of indeterminate worthiness of passing a test or progressing to the next level.

We shall return to the more criterion-referenced option later in the discussion but, for now, we must deal with the fact that the overwhelming majority of practitioners have *not* placed such an interpretation on the prescription of ATM. Two factors seem to have generated this situation: one deriving directly from Angoff's footnote to his first paragraph; the other deriving from, and perhaps concealed by, the subtle shift in language from that used by Angoff (1971) to that used widely today. Let us deal with the footnote first.

3. Angoff's Footnote Method (AFM)

As an exemplar of criterion referencing, Angoff's prescription stumbles at the first hurdle through the barely perceptible 'sleight-of-hand' that occurs as we switch from Angoff's first paragraph to its associated footnote.

Angoff (1971, p.515) offers an alternative to the procedure outlined in the first paragraph of ATM by specifying AFM as follows:

> A slight variation of this procedure is to ask each judge to state the *probability* that the "minimally acceptable person" would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who could answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score. A parallel procedure, of course, would be followed for the lowest honors score.

This footnote has achieved a prominence far outweighing its casual inclusion in the original article, specifying the most widely-used procedure for implementation of a method purported to produce a criterion-referenced standard that seeks to establish competence (Mills & Melican, 1988; Norcini, 2003; Amin, Chong, & Khoo, 2006; Lypson, Downing, Gruppen, & Yudkowsky, 2013). Nonetheless, AFM cannot be regarded as a 'slight variation' of ATM if ATM is implemented according to the strategy given in Section 2.1 A above. According to a strictly criterion-referenced view of 'minimally acceptable' for the establishment of a pass mark, each item in a test must be given a binary value of 0 or 1 as a multiplier of the respective mark attached to the item (1 for all 'must know' items, 0 for all other items). But this view cannot accommodate a compromise notion of 'probability' where the value of the probability may take non-binary values other than 0 or 1. The fact that Angoff (1971) regarded AFM as a 'slight variation' of ATM suggests that he may not have appreciated the extent to which he lost sight of his assumed criterion-referenced goal almost as soon as he tried to illustrate how it might be achieved.

While most applications of Angoff's methods have adopted the AFM approach of assigning probabilities over a continuous range between 0 or 1, judges have clearly found the procedure difficult (Norcini, 2003, p.466). Such difficulties apparently led to the 're-discovery' of ATM by Impara and Plake (1997), mistakenly reported by Jalili, Hejri, and Norcini (2011) as being proposed in 1997 as "another variation" of the Angoff method, now called the Yes/No Angoff method. In turn, difficulties with applying the Yes/No method to test items then led to the emergence of a "Three Layered Angoff" (TLA) method in which the ratings are Yes = 1, No = 0 and Maybe = 0.5 (Yudkowsky, Downing, & Popescu, 2008; Jalili *et al.*, 2011). The introduction of the "Maybe" category shows plainly that the focus has shifted from strict criterion referencing to some kind of norm referencing, the norm in this case being the examiner-conjured virtual image of a 'minimally acceptable' examinee. "Maybe" is not a category to which examiners should be in the habit of consigning significant chunks of their curriculum or batches of their questions for criterion referencing, but it is certainly a category that would be heavily populated by examiners attempting to predict the performance of students who cannot be judged with confidence as being of either pass-grade or fail-grade quality.

The intended criterion-referenced goal has been obscured by the attention given to the probability that an ill-defined subset of candidates could or would answer each item on a test successfully. Thus, it emerges that the AFM approach cannot be regarded as a 'slight variation' of the ATM approach, as presented by Angoff (1971), unless the ATM approach is *also* interpreted as a norm-referenced method, whereupon the ATM approach emerges as a particular extreme case of the more general AFM approach. Somewhere between the extreme binary approach of ATM and the widely-used continuum approach of AFM lies the more recent 'variation', the Yes/No/Maybe approach (Yudkowsky *et al.*, 2008; Jalili *et al.*, 2011). But, whichever of these three approaches has been used, it would appear that the original goal of achieving an absolute criterion-referenced standard has been obscured by the aforementioned subtle change in language to which we now turn.

3.1 Softening the language of ATM and AFM

While a central feature of Angoff's methods of standard setting is the definition of "minimally acceptable" persons, contemporary literature speaks, instead, of "borderline" persons. The difference between these labels reflects a shift from a sterner view of what is "minimally acceptable" to a more nebulous concept of what characterises a "borderline" student. This shift has blurred the conceptual distinction between "minimally acceptable" (more attuned to criterion referencing) and "borderline" (more attuned to norm referencing).

As discussed earlier, it is possible to put a criterion-referenced construction on the usage "minimally acceptable person" by defining such a person in terms of what knowledge, understanding and skills are required. By contrast, it does not seem possible to put such a construction on the concept of a 'borderline person'. On the contrary, the concept would appear from the outset to be *norm referenced*. Thus, the modern linguistic trend of substituting 'borderline' for "minimally acceptable" has facilitated and completed the confounding of norm referencing and criterion referencing in the standard-setting process.

4. Angoff's Second Paragraph

4.1 The Quest for Consensus

Angoff's (1971) second paragraph, reproduced above, describes how a consensus about the standards might be achieved among several independent "judges". This raises a number of procedural possibilities and issues according to the nature of the judging panel. Let us consider two extreme situations: A: a panel comprising judges having equal expertise in relation to all *n* test items; B: a panel comprising judges having expertise limited to different subsets of the *n* test items, such expertise showing overlap between individual judges and ranging from total overlap to zero overlap. In most medical schools running an integrated curriculum, it would be reasonable to expect that any given panel of judges will lie somewhere between these two extremes and, in practically all cases, much closer to the latter extreme.

A. Consensus among totally independent experts

In this idealised extreme case, each judge would be equally expert both on the entire content of the test and on the matter of assigning to each item an estimate for the proportion of borderline persons who would answer the item correctly (hereinafter called the "BL value" or, simply, the "BL").

However, real-world variability between judges might be expected to yield some slight variation in the BL estimates, resulting in some items being assigned different BL values by different judges, i.e., the creation of a subset of inconsistently estimated items *m*. This would indicate a need to achieve a consensus by sacrificing *post hoc* the absolute independence of the judges by averaging their estimates for each of the *m* items for which their estimates differed.

B. Consensus among inter-dependent partial experts

This situation is fraught with difficulty because the partial expertise is rarely manifest as complete expertise on a subset of *n* and zero expertise on the remainder. Rather, there will be a *gradient* of expertise for each judge, ranging from complete to zero among the *n* items. This difficulty could be overcome if each judge self-disqualified on each item for which less than complete expertise was possessed. However, in practice, consensus in these cases can only be reached among non-independent judges by having the less expert judges yield to the opinions expressed by the more expert judges on each item. While the intrusion of human nature will be an inescapable complication of this process, it is perhaps to be preferred over an alternative procedure in which each item would be assigned a BL according to an average of independently derived expert and inexpert inputs.

4.2 What is purported to happen

Nowadays, much emphasis is placed on training members of standard-setting panels in the art of grasping the concept of a 'borderline' student. For example, a typical description of the Angoff standard-setting process by Norcini (2003, p.465), under the heading *Angoff's method*, reads as follows:

> Judges are asked to first define the characteristics of a borderline group of examinees (a group with a 50% chance of passing). They then consider the difficulty and importance of the first item on the test. Each judge estimates what percentage of the hypothetical borderline examinees will respond correctly to the item. This judgement is often informed by data on the performance of the examinees. The judges discuss their estimates and are free to change them, and then proceed in the same manner through the remainder of the items on the test. The judges' estimates are averaged for each item and the cutpoint is set at the sum of these averages.

Despite the fact that this description is not without its problems, both logical and logistical, it is a fair description of what participants in contemporary standard-setting sessions in medical education imagine that they are doing. The problems are sufficiently important to warrant detailed dissection of this description, sentence by sentence.

> Judges are asked to first define the characteristics of a borderline group of examinees (a group with a 50% chance of passing).

This process of 'definition' is quite illusory. The word 'borderline' embodies the uncertainty attaching to persons who cannot be characterised as being clearly acceptable (deserving to pass) or clearly unacceptable (deserving to fail). Given this uncertainty, on what basis can such a group be said to have a 50% chance of passing? Only if the uncertainty of the examiners is symmetrical, i.e., if every conceivable type of person who is neither 'clearly acceptable' nor 'clearly unacceptable', is equally likely to have been wrongly excluded from either of these two categories.

Nonetheless, regardless of these issues, the focus is clearly on a subset of the cohort of examinees (whether actual, anticipated or imaginary), i.e., a norm-referenced focus, not a criterion-referenced focus. Moreover, given that judges are asked to conceive of such students as a hypothetical abstraction, based on their prior teaching and assessment experience, they can do nothing other than conjure up a norm-referenced concept (the 'borderline' person) and apply to it their own subjective guesswork. Given the difficulty in establishing a criterion-referenced absolute standard derived from such norm-referenced guesswork, it is not surprising that a desire has arisen among examiners to seek the protection and reassurance of group consensus among as large a collection of judges as can be mustered to define a standard for any given test.

> They then consider the difficulty and importance of the first item on the test.

This sentence directly and unnecessarily confounds the concepts of 'difficulty' and 'importance'. In fact, it is possible that these two concepts, in relation to individual test items, might be either essentially identical or totally unrelated. For example, the difficulty of an item might be judged to be a direct reflection of its importance, i.e., it is *important* for the item to be *difficult* as a property of its defining a minimally acceptable *criterion* of coping with difficulty. On the other hand, an item might be extremely *important* yet trivially *easy* for a correctly trained candidate, so that *importance* and *difficulty* are totally unrelated. Moreover, items that are unlikely to be answered successfully by any but the most exceptional candidates may possibly be very *difficult* yet essentially *unimportant*.

> Each judge estimates what percentage of the hypothetical borderline examinees will respond correctly to the item.

This is clearly a *norm-referenced* judgment, with no necessary link to any sense of *importance* of the item (*criterion referencing*). As already noted, the formation of such estimates is reported to be difficult (Lorge & Kruglov, 1953; Bejar, 1983; Impara & Plake, 1998; Norcini, 2003) although claims have been made that judges benefit from an iterative process whereby they can 'learn' from the estimates of their fellow judges formed during previous standardisation sessions dealing with the same assessment test (Cizek & Bunch, 2007, p.84).

However, except where special funding for educational research projects is available, iterative standard setting is beyond the resources and the practical exigencies of most medical schools. Moreover, when the test is a major examination of an integrated curriculum (reflecting current trends in medical education), it is not always possible to convene judging panels in which all specialities within the curriculum are adequately represented, let alone having multiple experts on each area capable of working out a consensus on the respective estimates.

At this point it is worth commending to the reader's attention the salutary study of AFM by Impara and Plake (1998) in which they "tested the ability of 26 classroom teachers to estimate item performance for two groups of their students on a locally developed district-wide science test." They found that "teachers' estimates of the average proportion correct" for 'borderline' students "were for the most part quite inaccurate", with only 23% of 1300 estimates focussed on these students being "accurate (defined as within .10 of actual item performance." Their concluding paragraph is worth quoting in full (Impara & Plake, 1998, p.80):

> The most salient conclusion we can draw from this study is that the use of a judgmental standard setting procedure that requires judges to estimate proportion-correct values, such as that proposed by Angoff (1971), may be questionable. The teachers in this study performed the estimation task in such a way that if their performance estimates were used to set a standard, the validity of the standard used to identify borderline students would be in question. *If teachers who have been with their students for most of the school year are unable to estimate student performance accurately using a test that is familiar to them, how can we expect other judges who may be less familiar with examinees to estimate item performance on a test those judges may never have seen before?* (emphasis added)

Returning to our dissection of Norcini's (2003, p.465) description, we find:

> This judgement is often informed by data on the performance of the examinees.

In our experience of standard setting, the practice of setting borderlines by reference to item statistics pertaining to past examinee cohorts ranges between two extremes:

A. *During a standard-setting session*, this practice is usually frowned upon as being *norm referenced*, the purpose of the standard setting being to establish a *criterion-referenced* pass mark. It is difficult at such sessions to establish acceptance of the fact that the participants are, in fact, being asked to predict a number that should be highly correlated with, and reasonably close to, the actual statistically derived proportion of LOW[4] students (as determined by post-test item analysis) that will be discovered to answer the item correctly. Attempts to establish the truth of this identity before the test is administered will often be contradicted by remarks such as, "No, that's *norm* referencing. We are *criterion* referencing." Such remarks are untrue, but they frequently win the day, presumably because the abovementioned confounding of *difficulty* and *importance* is fairly pervasive.

B. *During results review*, this practice might be encouraged if, as could be the case, reversion to *norm referencing* for a difficult item would *lower* the pass mark. For the record, this procedure has not been used at Gippsland Medical School.

While these extremes are mutually incompatible, they illustrate ways in which the standard-setting process can become compromised in practice.

[4] The 'LOW' subset of a cohort of examinees is the bottom 27% (approximately) as measured over the whole examination, including all multiple-choice questions and extended matching questions, but excluding short-answer questions (SAQs). The 'LOW' success rate for a given question is the proportion of the 'LOW' subset who answer the question correctly.

> The judges discuss their estimates and are free to change them, and then proceed in the same manner through the remainder of the items on the test.

This part of the process has already been considered above under section 4.1 B.

> The judges' estimates are averaged for each item and the cutpoint is set at the sum of these averages.

As already noted, the cutpoint may be subject to alteration when the results of the examination are reviewed.

4.2.1 What actually happens

The description here is suggested to be typical of what happens routinely in medical schools that apply the Angoff Footnote Method (AFM) of standard setting. It may not be typical of what happens in standard-setting sessions that form part of specially resourced educational research projects.

A standard-setting session is usually held several days after a draft of the examination paper has been pre-circulated among all academics involved in teaching and/or examining the unit. For most items, each BL has already been supplied by the respective item's author. In many cases, prior statistics deriving from item analysis are also attached to individual items that have been used in previous examinations.

Academics are encouraged to review the examination paper thoroughly, focussing particularly on their own areas of expertise, advising of any errors or recommendations for improvement, reviewing the BLs supplied and, where not supplied, suggesting BL scores. Academics who are unable to attend the standard-setting session are encouraged to submit their comments in writing so that they may be considered at the meeting.

At the meeting, attention is focussed only on those items that have been flagged for attention in the period prior to the meeting. It is noteworthy that concord between the BLs and their respective LOW success rates (where item statistics are available from prior examinations) is frequently absent. Where the disparity is severe, there may be a consensus at the meeting to alter the BL, but we have rarely seen any BL set lower than 0.3 prior to the examination, despite many LOW success rates being 0.2 or less (see Figure 1, lower right panel).

This disparity between BL predictions and actual LOW success rates finds resonance in the following quote from Norcini (2003, pp.465-466):

> Angoff's method is relatively easy to use, there is a sizeable body of research to support it, and it is frequently applied in licensing and certifying settings. It also has the virtue of focusing attention on each of the questions and thus can be very helpful from a test development perspective. This method produces absolute standards, so it is best suited to tests that seek to establish competence. However, judges sometimes feel as though there is no firm basis for their estimates and application of the method can be tiresome for longer tests.

While the standard-setting process (regardless of the method used) is certainly most helpful from a test development perspective, it cannot be supported that AFM produces absolute standards. Indeed, there is no firm basis for the estimates, as will now be discussed.

5. Angoff's Footnote Method in Action in a Small Medical School

Let us now turn to some assessment data obtained in three successive years from examinations set for first-year graduate-entry students at Monash University's Gippsland Medical School in 2010-2012. Prior to each examination, a standard-setting session was held in which a panel of interdependent partial experts was asked to reach a consensus, item by item, on estimating the proportion of borderline candidates who would answer each item correctly for multiple-choice questions (MCQs) and extended-matching questions (EMQs).[5]

5.1 Relation of BL estimates to different subsets of examinees

Figure 1 shows data plots relating the BL estimate associated with each MCQ or EMQ and the respective performance (P) success rate for the entire student cohort (P_{ALL}) and for the three subsets of candidates identified by statistical analysis of the results (P_{HIGH}, P_{MID} and P_{LOW}; see Footnote 4). All the data represented in Figure 1 were obtained from the 2010 cohort of 76 examinees answering 83 questions in the mid-semester 1 examination. Thus, although each data plot contains 83 data points, far fewer than this number are visible owing to superposition of many data points.

Figure 1 also shows the results of analysing the data using linear regression of P values on respective BL estimates. While there is no reason to expect uncomplicated linear correlations, it is reasonable to expect that the values of both the slope and R^2 of the linear regression would be greatest for the LOW subset and least for the HIGH subset. It is also to be expected, as found, that the ordinate intercept of the regression should be close to zero for the LOW subset and significantly greater than zero for the HIGH subset. Consistent with these expectations are the intermediate values of slope, R^2 and ordinate intercept observed for the MID subset and for the whole cohort (ALL).

Despite the fulfilment of these expectations, it is clear that the BL estimates arrived at through a consensual implementation of AFM are almost randomly and widely inaccurate; the performance success rates span much more than half the available range between 0 and 1 for almost all the BL estimate levels provided. The overall impression gleaned from these data is that academics' BL estimates are extremely poor predictors of examinee performance, and this impression is at its strongest in relation to the performance of the LOW subset.

[5] They were also asked to estimate the average mark likely to be obtained by a borderline candidate on each item for Short Answer Questions (SAQs), but this category of estimates is not included in the present analysis.

This impression of gross inaccuracy is maintained as we refine the focus to the 'Borderline'[6] subset as will be shown in the next subsection.

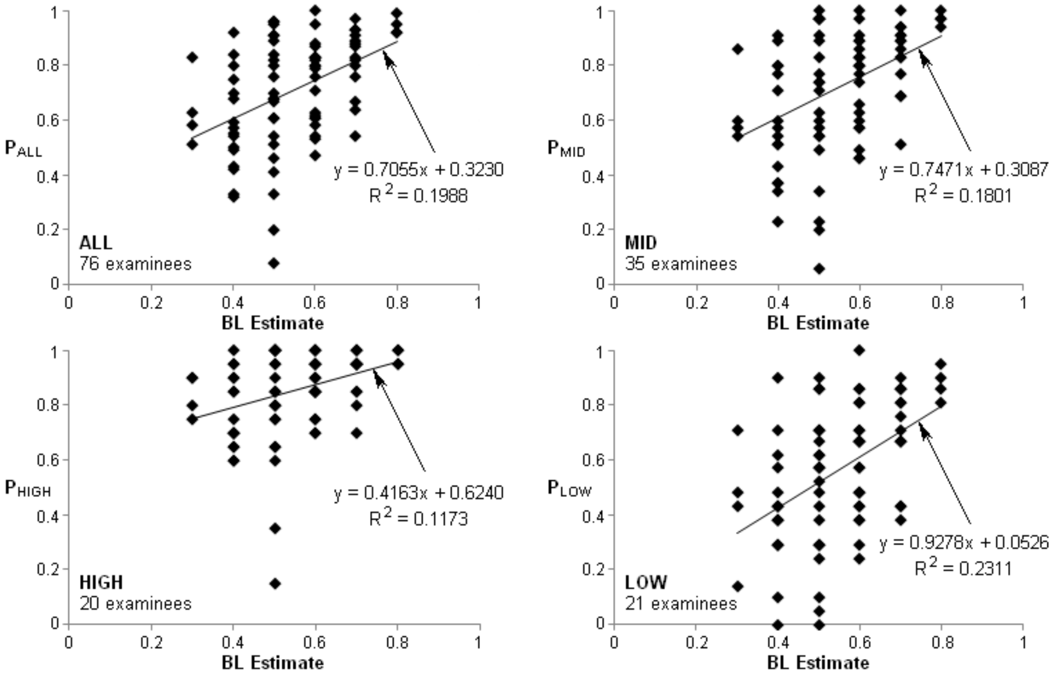

Figure 1: Data plots of performance rates, P, *vs* BL estimates for students answering 83 questions in a mid-semester 1 examination in 2010. The plots are for the entire cohort of 76 students (ALL, upper left), for the top 20 students (HIGH, lower left), for the 35 midrange students (MID, upper right) and the bottom 21 students (LOW, lower right).The linear trend line, linear regression equation and R^2 value are included on each respective plot.

5.2 Accuracy of BL estimates for the 'Borderline' subset of examinees

This 2010 examination was given again in 2011 and 2012 with essentially the same group of academics producing very similar styles of questions. The data plots relating the BL estimate associated with each MCQ or EMQ and the respective performance (P) success rate for the 'Borderline' subsets are shown in Figure 2 for eleven examinees in 2010 (upper left panel), three examinees in 2011 (upper right panel) and for nine examinees in 2012 (lower panel). For these particular data plots, it is highly appropriate to perform linear regression analysis on the relations between BL estimate and performance because the estimate is explicitly purported to predict the actual performance of the identified 'Borderline' students.

The persistently poor correlations already observed in the data plots of Figure 1 are also observed in the data shown in Figure 2, indicating that the academics' predictive ability did not improve when only the 'Borderline' data were included, and nor did it improve with experience. In all three years it was often

[6] 'Borderline' students are defined as examinees whose results fall within a range from one SEM above to two SEMs below the overall BL determined for the test, where the SEM is the standard error of measurement of examinees' scores.

found that 'Borderline' performance values spanning the entire possible range between 0 and 1 could be returned for groups of items assigned any given BL value by the examiners. Moreover, the highly respectable slope and ordinate intercept seen from the 2010 'Borderline' subset were not seen in the two following years.

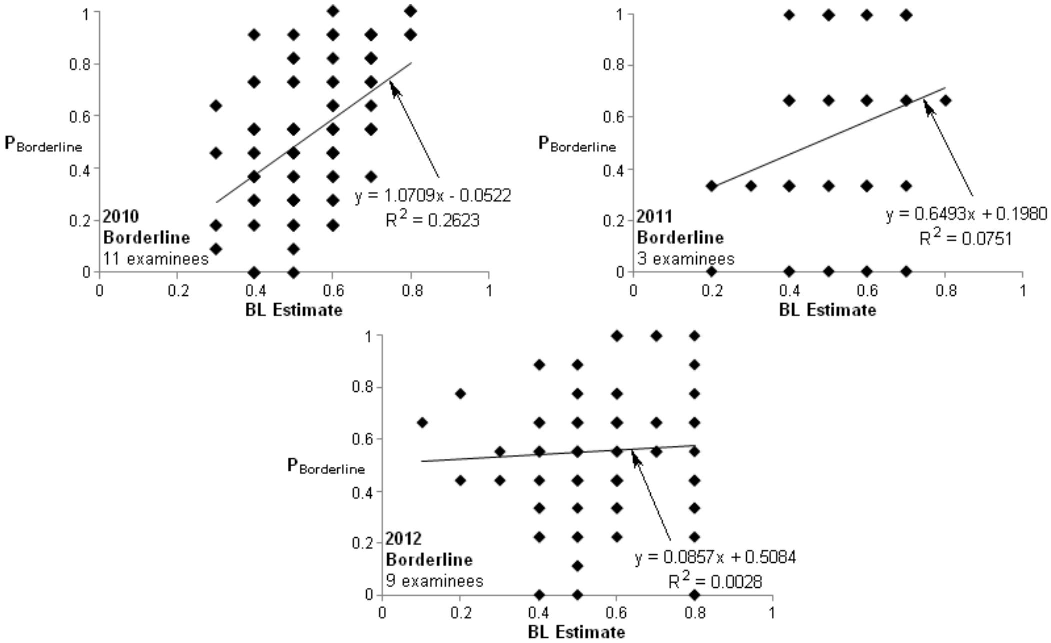

Figure 2: Data plots of performance rates, P *vs* BL estimates for 'Borderline' students answering questions in a mid-semester 1 examination in 2010 (83 questions, 11 examinees, upper left), in 2011 (82 questions, 3 examinees, upper right) and in 2012 (95 questions, 9 examinees, lower). The linear trend line, linear regression equation and R^2 value are included on each respective plot.

Figure 3 shows analysis of data gleaned from the same cohort as presented for the mid-semester 1 examination in 2012 (Figure 2 lower), showing analyses of data from the end-semester 1 (left) and mid-semester 2 (right) examinations from the same year. Interestingly, this cohort provided no data for such analysis in the 2012 end-semester 2 examination because all the students passed. That is, for the 2012 cohort, the numbers of 'Borderline' students identified by using Angoff's Footnote Method for the four successive mid-semester and end-semester examinations were 9, 9, 15 and 0, respectively. Any discussion or further exploration of this interesting finding would stray too far from the focus of the present critique and so will not be pursued here.

It might be objected that the small numbers of identified 'Borderline' examinees in the examination data reported here (ranging from 3 to 15 per examination, with one instance of zero) militate against drawing firm conclusions. However, the conclusions pertain to the accuracy of predictions of BL values for large numbers of questions, ranging from 75 to 95 per examination. Almost by definition, one hopes, the numbers of 'Borderline' candidates identified among cohorts of graduate-entry medical students should be small. The critique concerns the accuracy of the large number of predictions about individual questions that are made in relation to the actual performance on those questions by the small numbers of identified 'Borderline' candidates.

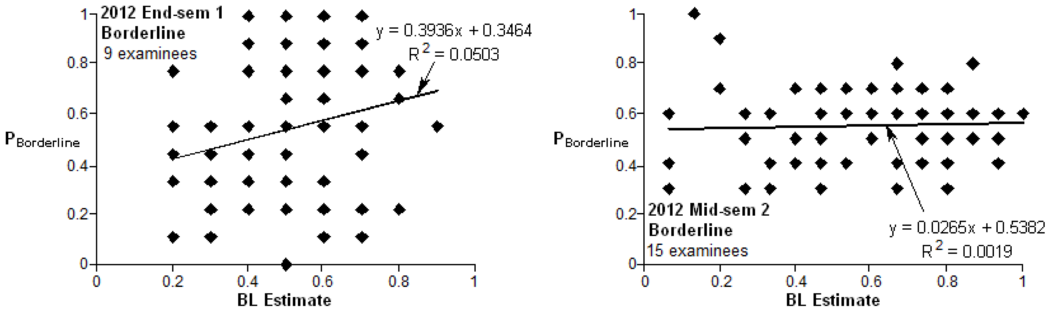

Figure 3: Data plots of performance rates, P *vs* BL estimates for 'Borderline' students answering questions in an end-semester 1 examination (87 questions, 9 examinees, left) and a mid-semester 2 examination (75 questions, 15 examinees, right) in 2012. The linear trend line, linear regression equation and R^2 value are included on each respective plot.

5.3 A want of feasibility and credibility for Angoff's Footnote Method

While this direct comparison of BL values with 'Borderline' performance values has been held to be a way of providing "useful checks on the passing score that is being chosen" (e.g., Kane, 1994, p.447), it seems likely that the poor predictive powers of academics makes the use of AFM an exercise in futility. The only comfort that can be taken from the data shown in Figures 2 and 3 is that the large errors of BL estimation are so randomly distributed that there may be no *systematic* error in the passing scores actually obtained by using AFM in this way. Thus the attempt to make accurate predictions, though futile and unsuccessful, may actually do little harm.

This record of poor prediction of BL values by academics involved in delivering an integrated medical curriculum should come as no surprise to anyone who has absorbed the results and conclusions of the study by Impara and Plake (1998) cited earlier. The least that can be suggested is that these BL predictions cannot be held up as exemplars of the goal of defining criterion-referenced absolute standards that are reliable and valid. On the contrary, the BL estimates of 'Borderline' examinees' performances shown in Figures 2 and 3 are visibly unreliable and invalid. In fact, the data shown in Figures 1 to 3 would seem, at worst, to vindicate Glass's (1978, p.259) forlorn conclusion that "setting performance standards on tests and exercises by known methods is a waste of time or worse" and, at best, to resonate with Impara and Plake's (1998) finding that even schoolteachers who are deeply familiar with a long-established test and their successive student cohorts are unable to predict the performance of 'borderline' students in such a way as to produce a reliable criterion-based standard.

As already noted, this study is not the first to question the feasibility and credibility of the Angoff process; nonetheless there seems to have been a well-established historical tendency for educators to proceed blithely ahead with their theories and methods without due acknowledgement of precedence or criticism.[7]

The data plotted in Figures 1 to 3 cast grave doubt on Norcini's (2003, p.466) claim that the Angoff (Footnote) Method produces absolute standards. The standards set for individual items do not seem to correlate in any reassuring way with the degree of *difficulty* of the items, as encountered by either the LOW subset of examinees or the TOTAL cohort. As for correlation with the *importance* of the items (cf. Norcini, 2003, p.465 and earlier discussion in Section 4.2), that would seem to be a matter totally beyond analysis.

The data reported here are consistent neither with the findings of Shepard (1994), who found that judges tend to overestimate low performers and underestimate high performers, nor with the opposite finding reported by Impara and Plake (1998, p.75) who found that teachers systematically underestimated the performance of 'borderline' students. It seems that examiners are as likely to underestimate as to overestimate performance success rates over a wide range of BL estimate values. These disparate findings would appear to provide further evidence that absolute standard setting by such prediction is unattainable.

6. A Possible Criterion-Referenced Implementation of ATM

Let us now consider how ATM might be interpreted and implemented using the more criterion-referenced construction of the concept of the "minimally acceptable person" suggested in Section 2.1 A. In such an approach there must be a strictly criterion-referenced focus on the *content* of the assessment items and not a norm-referenced focus on the performance probabilities of examinees. Let us consider a test comprising 100 items, each item carrying 1 mark, with no possibility of scoring fractional marks on any of the items. That is, for each item, a correct answer scores 1 while an incorrect answer scores zero. The method is to "go through the test item by item and decide whether" a "minimally acceptable person ... could answer correctly each item under consideration."

[7] Zieky (1995, p.10) records a disturbing fact about the landmark publications of Angoff (1971), Hills (1971) and Glaser and Nitko (1971), all appearing in the same 2nd Edition of *Educational Measurement*. They all failed to recognise the pioneering work of Nedelsky (1954) published in *Educational and Psychological Measurement*. Zieky (1995, p.6) notes that "concepts described by Nedelsky are found in more recent descriptions of methods of setting standards", and he finds that Nedelsky's concept of a 'borderline' student "corresponds to Angoff's (1971) 'minimally acceptable person', to Ebel's (1972) 'minimally qualified (barely passing) applicant,' and to the members of the 'borderline group' described by Zieky and Livingston (1977)." Despite the importance of Nedelsky's (1954) work, Zieky (1995, p.10) notes with interest that "neither Angoff nor Hills nor Glaser and Nitko referenced Nedelsky's article 'Absolute Grading Standards for Objective Tests.' The article was clearly relevant to the problems addressed in their chapters and had been printed in a major journal about 17 years earlier."

6.1 Unambiguous criterion referencing: focus on the test

By this more criterion-referenced definition, a “minimally acceptable person” *must* know, understand or accomplish certain well-defined facts, concepts or procedures, respectively. In other words, a “minimally acceptable person” *must* demonstrate *mastery* of certain well-defined *criteria*. Once the *criteria* have been defined according to the objectives of the curriculum, the determination of a criterion-referenced pass mark (or honours mark) becomes straightforward and unambiguous.

By thus substituting *must* for ‘could’ in ATM, the implication is that, of all the n items in a test, a given item, i, encapsulates *required material* if no person failing to show *mastery* of this material (i.e., failing to score 1 for item i) should be allowed to pass the test. The sum of marks attaching to all such *required items*, p, is therefore the pass mark defining the *mastery requirement* of the “minimally acceptable person”.

To follow Angoff’s suggestion that a similar procedure could be followed for the hypothetical lowest honours person, all that is required is that, in addition to the p ‘must know’ items already identified as required material for the “minimally acceptable person”, a further h ‘should know’ items be identified as required material for the lowest honours person. It then follows that, for a test containing a total of n items, there will be a number of items, $x = n - p - h$, that will be answered correctly only by exceptional candidates (‘nice to know’). This proposed distribution of the n test items is summarised in Table 1.

These considerations of *minimally acceptable* or *exceptional* performance can be applied equally to test items whether they encapsulate required knowledge, required understanding, required skills, or some combination of these attributes. It is important to note, therefore, that this method of standard setting is focussed entirely upon the test items insofar as they are identified as encapsulating *required* material at whatever level; the focus is not upon the examinee.

Table 1

Item type	Description
p	‘Must know/understand/accomplish’ – all items for which mastery is required by a minimally acceptable person.
h	‘Should know/understand/accomplish’ – all further items for which mastery is required by the lowest honours person.
x	‘Nice to know/understand/accomplish’ – these items are unlikely to be answered correctly by any but exceptional persons.
$n = p + h + x$	Total number of items (marks)

To construct such a test in which the pass mark is 50% and the honours mark is 85%, it is sufficient to ensure that p items carry 50% of the marks and h items carry 35% of the marks. In our example test, comprising 100 equally weighted items (1 mark per item), such a standard would be set by setting p = 50 and h = 35. But note that this implies that 50 p-category items, 35 h-category items and

15 *x*-category items have been *pre-identified independently* and brought together to produce such a combination of 100 items for the test so that such a standard obtains for the test. Other combinations of test items could be put together *prior to standardisation*, following which the standard setting would determine different values of *p*, *h* and *x*. These could then be assigned different marking weights so as to produce, if so desired, cutpoints at the 50% and 85% boundaries.

The above model is simply offered as a suggestion as to how a more consciously criterion-referenced approach to standard setting might be developed. It is not seen as a pure model; such a thing would seem to be unattainable. For example, it is possible for a candidate to score *p* marks while answering some of the *p*-category questions incorrectly, the balance of the marks coming from correct answers to questions in other categories. When one allows for the intrusion of guesswork into candidates' answers, including the unknowable proportions of informed and uninformed guesswork, the criterion-referenced goal of an absolute standard becomes even more illusory. Rather, this method of standard setting is offered as an approach to the development of tests that are explicitly tied to curriculum objectives and that allow for the capture of information about ranking of candidates in relation to the attainment of those objectives.

6.2 Teach to the objectives and test the objectives

Let us now turn to a simple observation, drawn from experience, that highlights the difficulty in believing that AFM can produce an absolute standard of any kind.

At Gippsland Medical School in the years 2008, 2009 and 2010, the electrophysiological material on propagation of the nerve action potential was delivered as a didactic lecture to first-year graduate-entry medical students. Among other things, the lecture dealt with the passive electrical properties (resistances, capacitances) of long cylindrical nerve axons and the effect of myelination on those properties. This topic is clinically important because of the disease state of multiple sclerosis in which nerve axons lose their myelin sheaths, leading to motor disability and death.

The lecture was always supported by comprehensive, detailed lecture notes made available online in two formats: as a linear text document and as an interactive 3-layered hypertext application. However, it was consistently found that students had difficulty in assimilating and applying the concepts underlying the effects of myelination or demyelination on electrical signalling in nerves. A typical multiple-choice question was set in 2010 as follows (correct answer in bold type):

What is the effect of myelination on a nerve fibre?

A. It increases the membrane resistance while reducing the membrane capacitance
B. It reduces the membrane resistance while increasing the membrane capacitance
C. It reduces both the membrane resistance and the membrane capacitance
D. It increases both the membrane resistance and the membrane capacitance

E. It has no effect on either the membrane resistance or the membrane capacitance, provided the influence of the electrogenic pump is ignored

The statistical item analysis for this question in 2010 was as follows:

```
ITEM  51: DIF=0.513, RPB=  0.478, CRPB=  0.427 (95% CON=  0.223, 0.595)
RBIS= 0.599, CRBIS= 0.535, IRI=0.239
   GROUP        N     INV    NF     OMIT   A*     B      C      D
TOTAL           76    0      0      0      0.51   0.36   0.08   0.05
          HIGH  20    0                    0.75   0.10   0.10   0.05
          MID   35    0                    0.60   0.37   0.00   0.03
          LOW   21    0                    0.14   0.57   0.19   0.10
   TEST SCORE MEAN %:                      76     65     66     65
   DISCRIMINATING POWER                    0.61   -0.47  -0.09  -0.05
          STANDARD ERROR OF D.P.           0.16   0.15   0.11   0.08
```

While it was regarded as disappointing that only 51% of the class answered the question correctly, this was consistent with observations in the preceding two years where students found this topic quite difficult. Note, however, that the question showed a high discriminating power of 0.61, with 75% of the HIGH group, 60% of the MID group and only 14% of the LOW group answering correctly. The underlined part of the analysis shows the performance of the LOW group of students (the 21 lowest performers on the overall examination out of a total cohort of 76 students).

In 2011 it was decided to replace the respective didactic lecture with a compulsory Tutorial for which students had to prepare answers to six set tasks. The six tasks were assigned for presentation among the cohort's six Problem-Based Learning (PBL) groups for cooperative preparation of a presentation to be posted online a few days before the tutorial. All students were required to prepare for the tutorial by studying all six of the online presentations. At the tutorial, students were selected at random to present their findings to the tutorial group (Presenters) or discuss the findings of other students (Discussants). In short, active learning was enforced in 2011, unlike in previous years where a didactic lecture was used. The same linear text document and interactive 3-layered hypertext application were used to support the tutorial as for the previous year's lecture. The relevant tutorial task was given as follows:

The Effect of Myelination on Passive Electrical Properties of Nerve Axons

a. Explain how myelin is formed in the central and peripheral nervous systems.
b. What are the principal features of a node of Ranvier?
c. How does myelination influence:
   - membrane resistance?
   - membrane capacitance?
d. Therefore, how does myelination affect the conduction velocity of a nerve fibre?

When the same MCQ was set in the 2011 examination it was given a BL score of 0.2 using AFM, based on the 2010 item analysis (i.e., LOW = 0.14). In the event, the item analysis in 2011 was as follows:

```
ITEM  11: DIF=0.841, RPB=  0.222, CRPB=  0.176 (95% CON= -0.034, 0.372)
RBIS= 0.334, CRBIS= 0.266, IRI=0.081
GROUP          N      INV     NF      OMIT   A*      B       C       D
TOTAL          88     0       0       0      0.84    0.09    0.02    0.05
        HIGH   25     0                      1.00    0.00    0.00    0.00
        MID    40     0                      0.77    0.10    0.05    0.08
        LOW    23     0                      0.78    0.17    0.00    0.04
  TEST SCORE MEAN %:                         70      62      66      67
  DISCRIMINATING POWER                       0.22    -0.17   0.00    -0.04
        STANDARD ERROR OF D.P.               0.09    0.08    0.00    0.04
```

Thus, for this particular question, there was a very large improvement in performance in 2011 relative to 2010 across the entire cohort, with 84% of the class answering the question correctly. Moreover, the question's discriminating power became much lower (0.22), with 100% of the HIGH group, 77% of the MID group and 78% of the LOW group answering correctly. The underlined part of the analysis shows the performance of the LOW group of students (the 23 lowest performers on the overall examination out of a total cohort of 88 students).

When the same question was run in the corresponding 2012 examination after delivering the material using the same 2011 active learning model, its performance statistics were as follows:

```
ITEM  12: DIF=0.721, RPB=  0.216, CRPB=  0.161 (95% CON= -0.053, 0.360)
RBIS= 0.289, CRBIS= 0.214, IRI=0.097
  GROUP        N      INV     NF      OMIT   A*      B       C       D       E
TOTAL          86     0       0       0      0.72    0.14    0.03    0.09    0.01
        HIGH   24     0                      0.83    0.13    0.00    0.04    0.00
        MID    38     0                      0.74    0.13    0.00    0.13    0.00
        LOW    24     0                      0.58    0.17    0.13    0.08    0.04
  TEST SCORE MEAN %:                         69      66      58      68      60
  DISCRIMINATING POWER                       0.25    -0.04   -0.13   -0.04   -0.04
        STANDARD ERROR OF D.P.               0.13    0.10    0.07    0.07    0.04
```

This represents a slight drop in performance in 2012 relative to 2011, but still much improved relative to 2010 across the entire cohort, with 72% of the class answering the question correctly. The question's discriminating power increased slightly (0.25), with 83% of the HIGH group, 74% of the MID group and 58% of the LOW group answering correctly. The underlined part of the analysis shows the performance of the LOW group of students (the 24 lowest performers on the overall examination out of a total cohort of 86 students).

These observations on the performance statistics of a single question show that the BL estimations cannot possibly produce the claimed absolute standard of criterion referencing, and this remains true however the observations are interpreted. They could have been due in part to having superior cohorts of students in 2011 and 2012 relative to previous years, and it must be acknowledged that the level of competition and the cut-off scores associated with gaining admission as graduate-entry medical students are increasing year by year. However, it is strongly suggested here that most of the differences are due to the substitution of active learning for what, in the past, had been largely passive learning; this is hardly a surprising result. Whatever the relative contributions of these two causes to these observations might be, the fact remains that the claim (Norcini, 2003, pp.465-466) of setting absolute standards in BL predictions using the Angoff Footnote Method is entirely without foundation. The performance of 'borderline' students is more dependent on the methods of teaching and learning applied than it is on the intrinsic difficulty of the content.

This result certainly accords with the conclusion of Glass (1978, p.239), who wrote:

> The vagaries of teaching and measurement are so poorly understood that the a priori statement of performance standards is foolhardy.

However, the suggested remedy for these problems is not to seek some unattainable absolute standard, but to apply criterion-referenced (i.e., curriculum-determined) standards of relative importance to test items, ensuring that all items test the objectives, and then teach to the objectives.

6.3 Constructing a criterion-referenced standard for a test

The following checklist of requirements is suggested for producing a curriculum-determined standard according to the criterion-referenced interpretation of ATM in which a test comprises a mixture of items in the *p*, *h*, and *x* categories described above in Section 6.1 and Table 1:

- Avoid undue dependence on centralised question banks; there can be no certainty that the questions have been composed in relation to the currently operative learning objectives, or with a view to accommodating the problems of standard setting addressed in this paper. Some such questions may prove useful, but only after they have been subjected to careful scrutiny to decide how they might be assigned among the *p*, *h* and *x* categories.

- Before presenting the formal teaching/learning opportunity (lecture, tutorial, practical class and any supporting handouts, online documentation, etc.), identify the respective learning objectives and construct at least one item in each category to address each learning objective as follows:

    - **Category *p*** – *Must know* or *Must understand* or *Must accomplish* – A person who fails to answer such questions correctly is not yet ready to proceed to the next year level. These questions cover essential information, understanding and skills. Such material is not necessarily easy, although in many cases it may be.

  - **Category *h*** – *Should know* or *Should understand* or *Should accomplish* – A person who fails to answer such questions correctly is unworthy of attaining a *Credit*, but such failure is not a significant impediment to progression to the next year level. These questions may be more challenging than those in the preceding category, but should not venture beyond material that has been presented in teaching/learning sessions or documented in online course materials.

  - **Category *x*** – *Nice to know* or *Nice to understand* or *Nice to accomplish* – A person who answers one quarter of such questions correctly is worthy of attaining a *Distinction*; a person who answers half of such questions correctly is worthy of attaining a *High Distinction*. These questions should be relevant to the material covered in the respective session but not necessarily covered in detail or even directly mentioned at all. They could explore material that a good student might be expected to pick up through further self-directed study.

- When presenting the formal teaching/learning opportunity and preparing any supporting online documentation, it is important to *teach to the objectives* with respect to the *p*-category and *h*-category items. This underlines the importance of identifying the learning objectives and preparing the targeted test items *before* preparing the associated teaching and learning resources.

- As with conventional application of the various Angovian methods, examiners should have the opportunity to submit their questions to colleagues for feedback on their individual standard-setting judgments (in this case, the distribution into *p*, *h* and *x* categories). In particular, there may be need for review of the assignment of questions among the three categories both prior to, and following, the administration of the test. This would provide essential information regarding the accuracy and feasibility of such attempts to set relative standards unambiguously.

- When the test results are known, review the allocation of students among the various grades and, if the results of applying the implicit ATM-derived criterion-referenced standards are unacceptable, then let the allocation of grades be influenced by norm-referenced considerations. If too much norm-referenced adjustment has to be made to the ATM-derived standard, then use that information to guide an inquiry into the construction and allocation of questions among the three categories, the teaching and learning resources provided, and the communication of information to students.

7 'Borderline' Students in Graduate-Entry Medical Schools

Competition to gain admission to graduate-entry medical courses is very strong in Australia. Those who succeed do so by obtaining increasingly high marks in the Graduate Australian Medical School Admissions Test (GAMSAT) examination. Given that such students have already demonstrated academic success at tertiary level, that they remain sufficiently motivated to study medicine and that, at Gippsland Medical School and many other schools, they are further assessed at interview, there is good reason to suppose that every student gaining admission to graduate-entry medicine should be able to proceed to graduation in the minimum time. That is, we should not expect to find more than a handful of 'borderline' students in each cohort and we should not be surprised occasionally to find none at all. On the contrary, such expectations flow naturally from the confidence we place in the selection process for graduate-entry medicine. The reasonable default expectation is that all graduate-entry medical students should pass.

It follows that the existence of 'borderline' students in graduate-entry medical cohorts simply demonstrates that the selection process is imperfect and unable to guarantee an absolute standard. Norm referencing will find such students out, as it always has. It would seem unrealistic and unproductive to search for an absolute objective standard such as would relieve examiners of the need to take responsibility for exercising subjective judgments, when required, to deal with 'borderline' students. As Kane (1994, p.427) observed:

> We create the standard; there is no gold standard for us to find, and the choices we make about where to set the standard are matters of judgment.

8 Conclusion

As this discussion has been more critical than supportive of existing applications of Angoff's methods and their derivatives, it is important to clarify that this was never intended to be a direct criticism of Angoff (1971)[8]. Rather, it has been a discussion of the possibility that succeeding generations of educationists may have been too uncritical in their application of a method that was originally offered quite casually as a brief incidental insertion within a very large chapter devoted to other assessment issues.

It is suggested that:

- Generations of educators who have sought to implement Angovian methods for defining criterion-referenced standards have, whether consciously or not, been guided by a belief that an objective absolute standard is achievable.

- This belief has been undermined from the outset by:

[8] Angoff is even reported to have claimed in the early 1980s that the true originator of the method attributed to him was the American mathematician Ledyard Tucker, as recorded by Jaeger (1989, p.493) and Zieky (1995, p. 9 footnote 3).

  - blurring the focus on criterion referencing with an insufficiently precise definition of a "minimally acceptable person", and

  - replacement of the already ill-defined concept of a "minimally acceptable person" with the norm-referenced concept of a 'borderline' person.

- Rather than Angoff's Footnote Method (AFM) being a "slight variation" of Angoff's Text Method (ATM), the dominant interpretation and implementation of Angovian methods of the past forty years reveal ATM to be an extreme, binary example of the more generally used probability continuum of AFM, all such applications being norm referenced by focussing on the examinee rather than on the curriculum.

- Unless standard setting takes place in the context of a generously funded educational research project, the search for consensus among panels of independent experts is not feasible in the routine management of assessment of an integrated curriculum in a graduate-entry medical school.

- Academics' predictions of 'borderline' student performance are manifestly inaccurate in a small rural medical school and are unlikely anywhere to be more accurate than the documented inaccuracy reported among secondary school science educators (Impara and Plake, 1998). However, the prediction errors, though wide-ranging, are possibly sufficiently random to generate no serious systematic error in the performance estimates for 'borderline' students averaged over a test comprising many assessment items. Thus, the inaccuracy of the predictions may not do significant harm, even though the expenditure of resources in generating them may not be justified.

- A more rigorously prescriptive interpretation of ATM raises the possibility of applying criterion-referenced (i.e., curriculum-determined) standards directly to test items, with items distributed into 'must know', 'should know' and 'nice to know' categories.

- While standard setting should be done as objectively as possible, the intrusion of imperfection in student selection procedures and assessment procedures will always require examiners to take responsibility for exercising subjective judgments.

Acknowledgements

Thanks are due to Professor William Hart and Associate Professor Elmer Villanueva for their helpful comments on earlier drafts of this paper.

References

Amin, Z., Chong, Y.S., & Khoo H.E. (2006). *Practical Guide to Medical Student Assessment*. London, England: World Scientific Publishing Co.

Angoff, W. H. (1971). Scales, Norms, and Equivalent Scores. In R. L. Thorndike (Ed.), *Educational Measurement*, 2nd edn, (pp. 508-600). Washington, DC: American Council on Education.

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, 7(3), 303-310.

Cizek, G .J., & Bunch, M.B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards for Tests*. Thousand Oaks, CA: Sage Publications.

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.

Ebel, R. L. (1972). *Essentials of educational measurement*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.

Glaser, R., & Nitko, A. J. (1971). Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational Measurement*, 2nd edn, (pp. 625-670). Washington, DC: American Council on Education.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15(4), 237-261.

Hills, J. R. (1971). Use of measurement in selection and placement. In R. L. Thorndike (Ed.), *Educational Measurement*, 2nd edn, (pp. 680-732). Washington, DC: American Council on Education.

Impara, J. C., & Plake, B. S. (1997). Standard Setting: An Alternative Approach. *Journal of Educational Measurement*, 34(4), 353-366.

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.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement*, 3rdedn, (pp. 485-514). New York, NY: American Council on Education/Macmillan.

Jalili, M., Hejri, S. M., & Norcini, J. J. (2011). Comparison of two methods of standard setting: the performance of the three-level Angoff method. *Medical Education*, 45(12), 1199-1208.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425-461.

Lorge, I., & Kruglov, L.K. (1953). The improvement of the estimates of test difficulty. *Educational and Psychological Measurement*, 13(1), 34-46.

Lypson, M. L., Downing, S. M., Gruppen, L. D., & Yudkowsky, R. (2013). Applying the Bookmark method to medical education: Standard setting for an aseptic technique station. *Medical Teacher*, 35, 581-585.

Mills, C. N., & Melican, G. J. (1988). Estimating and adjusting cut off scores: Features of selected methods. *Applied Measurement in Education*, 1(3), 261-275.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14(1), 3-19.

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

Searle, J. (2000). Defining competency – the role of standard setting. *Medical Education*, 34(5), 363-366.

Shepard, L. A. (1994, October). Implications for standard setting of the NAE evaluation of NAEP achievement levels. Paper presented at the Joint Conference on Standard Setting for Large Scale Assessments, National Assessment Governing Board. Washington, DC: National Center for Educational Statistics.

Williams, P. (2008). Assessing context-based learning: not only rigorous but also relevant. *Assessment & Evaluation in Higher Education*, 33(4), 395-408.

Yudkowsky, R., Downing, S. M., & Popescu, M. (2008). Setting standards for performance tests: a pilot study of a three-level Angoff method. *Academic Medicine*, 83(10), S13-S16.

Zieky, M. J. (1995). A historical perspective on setting standards. In *Proceedings of the Joint Conference on Standard Setting for Large Scale Assessments* (pp. 1-38). Washington, DC: National Assessment Governing Board and National Center for Educational Statistics.

Zieky, M. J. and Livingston, S. A. (1977). *Manual for setting standards on the Basic Skills Assessment tests*. Princeton, NJ: Educational Testing Service.