Journal of Contemporary Medical Education
www.scopemed.org
DOI: 10.5455/jcme.20151019104847

Review Article

Fortune-tellers or content specialists: challenging the standard setting paradigm in medical education programmes

Margaret MacDougall[1], Gregory E Stone[2]

[1]Centre for Population Health Sciences, College of Medicine and Veterinary Medicine, Teviot Place, Edinburgh, Midlothian, EH8 9AG, United Kingdom.
[2]Department of Educational Foundations and Leadership, Judith Herb College of Education, University of Toledo, MS#921, Toledo, OH 43606.

Address for correspondence:
Margaret MacDougall, Centre for Population Health Sciences, College of Medicine and Veterinary Medicine, Teviot Place, Edinburgh, Midlothian, EH8 9AG, United Kingdom.
Margaret.MacDougall@ed.ac.uk

Received: November 27, 2014
Accepted: September 12, 2015
Published: November 05, 2015

ABSTRACT

The veracity of Objective Standard Setting (OSS) as a modern approach to criterion-referenced standard setting has been reported for healthcare student assessment in the USA, while in other countries, OSS remains unrecognized. OSS upholds the foundational principle for itemized tests that judges should base their decisions on test item content. Moreover, it presents judges with a conceptually transparent decision procedure. This contrasts with the predictions concerning a hypothetical borderline candidate which typify Angoff procedures. Furthermore, the iterative process involved in the Angoff standard setting task incurs financial and administrative burdens, thus creating the potential to cut corners through recruiting fewer judges. The underlying objective of homogenizing the test standard undermines its validity, while circumventing reputable standard setting principles. While the Rasch model offers an objective approach to predicting successful outcomes, combining Rasch and Angoff procedures does not resolve the validity problem for Angoff-based pass marks. This commentary highlights the virtues of OSS relative to the modified Angoff approach in the standard setting of itemized tests. It also identifies gaps in the research literature that should be addressed to strengthen the case for using OSS on an international scale for high-stakes assessments within healthcare disciplines as a testing ground for other disciplines.

KEY WORDS: Angoff, Criterion-referenced assessment, Medical education, Objective standard setting, Rasch model, Validity

INTRODUCTION

The practise of criterion-referenced standard setting has remained relatively stagnant for decades. While a number of models have been proposed, as has been recognized elsewhere, [1-4] each model tends to fit into one of two schools of thought. One set of standard setting models focuses on predictions of test-taker success on items, based, theoretically, on the difficulty and content presented within the item. Included in this set are models defined by Ebel, Hofstee, Nedelsky, Jaeger, and most notably, Angoff, among others. A second set of alternative models focuses directly on content, and the portion of that content necessary to claim practical test-taker mastery. These models include both the Objective Standard Setting Model (OSS) and Bookmarking/Mapmarking. While the Angoff model is arguably the most popular and well-known model in current use, alternatives exist which have helped in improving practise and addressing limitations associated with their predecessors. The newer models are distinctly different from their traditional counterparts, and while evidence has been widely presented in North America for their usefulness, they have yet to reach a global audience, or be evaluated in a range of assessment settings. The goal of this paper is to rectify this oversight, and present for broad consumption a discussion of the benefits and limitations of the two sets of models. Tradition will be represented by the Angoff model, while Objective Standard Setting (OSS) will be offered as a viable alternative.

The original Angoff method reported by Angoff in 1971 [5] required judges to decide separately for each item on a presented examination whether hypothetical, minimally competent examinees (MCEs) could answer that item correctly. A 1 or 0 was assigned to each item according to whether or not an individual judge anticipated that an MCE should be able to answer the item correctly. The final cut-off was determined through aggregating the resultant scores for individual judges across items and then taking a final average across judges. This approach, which relied more fundamentally on content, was almost immediately revised through a process which has come to be known as a *modified* Angoff approach.

The key standard setting task which typifies a *modified* Angoff procedure originates from a suggestion made in a footnote to Angoff's description of the above method. In this modification, judges are required to make item-specific judgements regarding the probability of a hypothetical MCE providing a correct or satisfactory response. Just as with the scores in the original Angoff procedure, speculated proportions are then summed across items for each judge and in turn combined to form an average across judges. This more usual interpretation of the term 'modified Angoff'

ought to be carefully distinguished from that assumed by Senthong et al. In the latter case, the term refers to a procedure involving the original Angoff standard setting task but with the pass mark adjusted according to both level of concordance between judges and the standard error of measurement (SEM). [6] This interpretation will not be assumed in this paper.

The modified Angoff standard setting task involves a change of focus from content to MCE performance, and the reaching of a consensus on attributes and behaviours which fall under the concept of MCE. At the initial stages, this might involve a brainstorming exercise whereby judges volunteer characteristic descriptors of the MCE. Examples include "hesitant", "unsure", "slightly disorganised", "covers most of the important things", "a bit awkward in communication", "has OKish clinical skills", "patchy knowledge" or "safe". (Dr K A Boursicot, St George's, University of London Advanced Course in Medical Education Assessment, 2007) Multiple iterations of this step frequently ensue, involving the use of normative performance data for existing examinees or for a previous examinee group. In the latter case, the potential impact of initial judge decisions (consequential data) might be compared with corresponding actual results and associated item difficulty ratings, including p values, which, in test item analysis, correspond to the percentages of examinees who responded correctly to the given test items. [7]

Typically, statistical data, including inter-judge consistency measures, frequency distributions for judge estimates, and individual judge cut-scores, are also included at this stage to allow judges to evaluate their initial judgements and calibrate their estimates accordingly. As such, the iterations are designed to produce consensus across judges. Indeed, as a credential for evaluating the modified Angoff procedure, in a published Association for Medical Education in Europe (AMEE) guide on standard setting practise, it is recommended that

"*Evaluation materials should include data on the first and second ratings of the panellists for each of the test components rated, which should demonstrate increased consensus of raters.*" [8]

Thus, it is implicit that scores across judges are expected to converge towards a suitable cut-off for a hypothetical MCE.

In practise, applications of modified Angoff procedures allow judges to discuss and adjust their initial estimates. Furthermore, prior to administration of the above tasks, there is frequently an important preliminary stage wherein, during an initial orientation relating to the nature and purpose of the test, judges define what characterizes a MCE. This stage is important, given Angoff's original recommendation that judges '[keep] the hypothetical "minimally acceptable person" in mind' [9] while making their decisions.

Relevant discussions may, for example, include speculating whether a MCE is the type of candidate who has a 50% chance of passing. Here, some judges might focus on the likely performance of one or more students whose performances they have observed in a previous test setting, while others may take a more abstract approach. However, the Angoff standard setting task, if properly understood, should involve a *randomly* selected hypothetical MCE. [10] This point is particularly pertinent given the intrinsic need for human cognition to grapple with infinity and the inherent difficulties therein! Thus, an emphasis is placed on judge ability to make speculative predictions in an infinite space, that is, impossible predictions. This is to the neglect of the expertise of the judge concerning the discipline-specific content of individual test items. Equivalently, the judge is placed in the ill-fitting shoes of a fortune-teller to the neglect of their invaluable expertise as a content specialist. These observations are critical, as they lend weight to concerns raised previously that the Angoff procedure is "fundamentally flawed" because it depends on cognitive judgements of probability that are "nearly impossible" to make. [11, 12]

Findings in psychology reveal "that dichotomous processing is a fundamental phenomenon of the human mind; and requiring a dichotomous response … is the most effective way to collect responses from people" [12]. Contrasting the main *unmodified* Angoff standard setting task, which requires simpler yes/no decisions, to that involved in the *modified* approach, which requires predictions of probability [13], it is clear that the accuracy and clarity of the modified Angoff approach to standard setting must be questioned.

Indeed, it has been recognized for several decades that, even among the statistically well-versed, the prediction of probabilities under conditions of uncertainty is typically biased by human propensity to rely on heuristics. Within such contexts, too much weight is attributed to irrelevant or less relevant types of evidence on the grounds that they appear intuitive [14]. Furthermore, it is legitimate to question the capacity of judges to retain the concept of *hypothetical* MCE throughout the steps of a modified Angoff standard setting process and hence the reliability of the pass mark arising from this process. [15]

Moreover, contrary to what De Champlain suggests, [4] the key task of predicting proportions of MCEs who ought to pass individual items is in conflict with the original intention of the early pioneers of standard setting, Nedelsky, Angoff and Ebel. In presenting criterion-referenced assessment, it is clear that they each separately required that the criterion for meeting a test standard should be grounded on test content, not examinee performance. [5, 16, 17] In particular, with respect to standard setting, Nedelsky observed that,

"*It is the essence of the proposed technique that the standard of achievement is arrived at by a detailed consideration of individual items of the test.*" [17]

In reference to deciding the pass score for his 'hypothetical "minimally acceptable person"', and prior to any modifications, Angoff suggested working "through the test item by item [to] decide whether such a person could answer correctly each item under consideration." [5] Similarly, Ebel highlighted the idea of defining the "passing score … on the basis of the pooled judgements of experts on the relevance and difficulty of each item in the test." [16] All of these original recommendations are strongly focused on item content.

Similar remarks apply to the many variants of the (standard) modified Angoff approach, as presented above, which have evolved over the last four decades. The distinguishing features of these variants depend on whether or not:

- a definition of minimal competence is provided for judges at the outset or constructed by them as a group for use in their individual decisions;
- at an early stage of the algorithm, judges are at liberty to use their personal notions of minimal competence without the need for prior consensus;
- judges make their decisions in a separate location from the standard setting meeting;
- judges sit the exam themselves

  and/or

- correct answers are withheld from judges.

The standard modified Angoff procedure and its variants remain among the more popular approaches in medical education examinations today. For example, in UK undergraduate medicine, the majority of written assessments involve multiple choice and short answer questions, with the standard being set using an Angoff approach. Cizek, Kane and Plake defend the Angoff approach (in its various forms) in terms of popularity as "widely accepted and praised" [18], as having been used "on a host of licensure and certification tests, as well as on numerous state testing programs, without major complaints from the judges involved" [19] and as probably "the most popular standard setting method in use today" [10]. Their viewpoints are, however, a natural product of the standard setting culture of the late 1980s wherein traditional approaches were established and research boundaries largely confined to finding variants of these approaches. It is this type of culture that may have contributed to the finding recently noted by McManus et al. that "most validation of judgemental methods such as Angoff rely for their validation mainly on repeated assertion of validity of process rather than any formal demonstration." [20] This type of defect, if left uncorrected, can leave medical schools exposed to legal and political threats arising from student complaints. [21]

The Angoff methodology, where this involves the modified task defined above, is deceptively simple to explain, which may lend appeal when choosing Angoff approaches. It is less clear, however, that the underlying tasks are achievable or even appropriate.

In using the term "modified Angoff" in future sections of this paper, we will focus on the standard modified Angoff procedure as defined above as a paradigm for instigating change in arriving at a more valid pass mark. This is under the recognition that many of the observations made by way of comparison of this particular procedure with the more objective methodologies to be presented here also apply with similar modified Angoff approaches, including those forthcoming from the above list of distinguishing features. With this clarification in place, the weaknesses of the modified Angoff approach provide a sound basis for defending a more recent Rasch model-dependent approach, known as *Objective Standard Setting* (OSS), as a suitable alternative in medical education. Through appeal to underlying methodology and experimental findings, this paper will highlight the virtues of OSS relative to the modified Angoff approach. As such, this work should provide some inspiration for a fresh perspective on research into standard setting methods in medical education while, as we shall see, upholding the foundational principles at the heart of criterion-referenced assessment.

VIRTUES OF THE RASCH MODEL

Within the context of criterion-referenced assessment, where scores are mainly absolute rather than relative, the status of the Rasch model in supporting an objective approach to standard setting is recognized. In medical education assessment, the corresponding criterion may be thought of as representing a particular level on a scale of achievement for a construct, such as *clinical competence or eligibility for certification in a clinical specialty or mastery*.

The Rasch model belongs to the more general family of Item Response Theory (IRT) models, which, in assessment contexts, are characterized by their capacity to estimate the probability of a discrete response to an item based on both examinee and item parameters. In the case of the original (or, basic) Rasch model, [22] these parameters are exclusively examinee ability and item difficulty, there being one item parameter. For dichotomously scored items, the defining equation for this Rasch model is,

$$P_i(\theta) = P_i(X_i = 1 \mid \theta) = \frac{1}{\{1+\exp[-(\theta - b_i)]\}}, \qquad (1)$$

where i ranges over all test items; 'θ', 'b_i' and 'P_i' denote the ability level for an examinee, difficulty level of item i, and item response probability function for a correct response to item i, respectively; the symbol '$\mid$' is used to indicate that the value of X_i is conditional on that of θ, and X_i is a dichotomous variable assuming the values 1or 0 according as to whether a

correct or incorrect response, respectively is obtained for item *i*. While the Rasch model can be generalized in various ways to include polytomously scored items, [23] the basic model is represented above for simplicity. From equation (1), it can be seen that, where all items are dichotomously scored, the item characteristic function for item *i*, obtained from the sum of the products of P_i and the corresponding response score for item *i*, is equivalent to P_i. For any given test item, the corresponding item characteristic curve (Fig. 1) displays the item response probability for a correct response against examinee ability.

A representative experimental group of examinees is selected. For high-stakes cases, such as 'finals' in undergraduate Medicine, the minimum sample size for this group should typically be 250. [24] This is to optimize stability and accuracy of item parameters and ultimately, of the derived pass mark across similar samples of examinees. Given performance data across all items for the experimental sample, ability and difficulty ratings are conveniently estimated on the same log-odds (or, 'logit') scale. The ability ratings of existing candidates are typically estimated from the resultant Rasch model using maximum likelihood estimation or Bayesian methods. [25] According to equation (1), for the Rasch model, the probability of a candidate providing the correct response for a particular item is 0.50 if and only if examinee ability matches item difficulty. Flatter ICCs correspond to more difficult items, as higher abilities levels are required to achieve correct responses with specific probabilities. This is illustrated for items 1 and 2 in Fig. 1, where for a correct response probability of 0.5, the flatter curve for item 2 maps this probability to a higher ability level of 1.5 compared to that of -0.5 with the steeper curve for item 1.

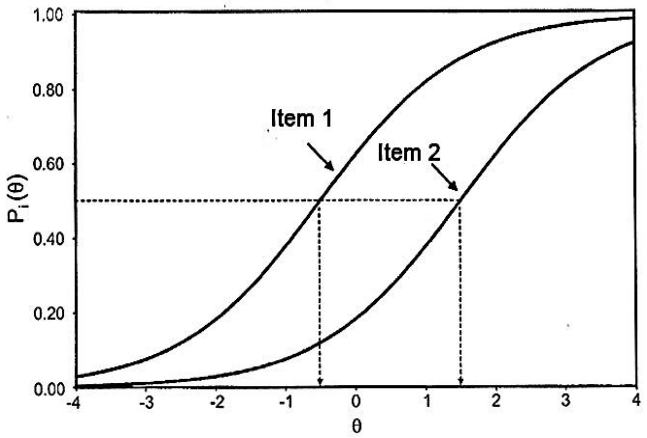

Figure 1. Item characteristic curves for two Rasch model items. θ and P_i denote the ability level for an examinee and item response probability function for a correct response to item *i* (*i* = 1, 2, respectively). Dotted lines correspond to derivation of ability levels for item response probability of 0.5
Adapted from Figure 4.1 of Yen & Fitzpatrick (2006).

Typically, it is these examinee abilities which are used to determine the pass mark (or, cut-score) as the expected raw score given the ability score in logits. This is accomplished as part of the Rasch analysis procedure using a test characteristic function which, for a given ability level, is expressed as the sum of the individual item characteristic functions across test items and examinees, or from the corresponding test characteristic curve. [25, 26]

A key benefit of the Rasch model over and above other forms of IRT models is that of *parameter separation*. In particular, unlike the item difficulty estimates of classical test theory, Rasch model item difficulty ratings are determined independently of the ability ratings of the individual examinees [25, 26] – a property known as *specific objectivity* [27]. Thus, the difficulty ratings forthcoming from item calibration are generalizable for use with future examinee cohorts. Also, at the level of the individual examinee, the ability rating is estimated independently of the difficulty ratings of the test items. Furthermore, for tests with some common items from one year to another, specific objectivity has a further benefit to offer. Once the standard from the previous test is conveniently transferred to the scale of the later test through Rasch calibration [28], specific objectivity ensures that the standard of the test is maintained irrespective of fluctuations in examinee abilities between individual years.

RASCH ANALYSIS IN CONTEMPORARY MEDICAL EDUCATION

While "Rasch analysis has been applied widely in medical assessment" [29], within the specific context of medical education, the tendency has been to use the Rasch model retrospectively to monitor quality of assessment. Examples include use of item fit analysis to test for variability in item difficulty across schools or styles of test administration [29, 30] and for evidence of unidimensionality in the scores forthcoming from a test of clinical competence. [31] Further examples include testing for rater leniency or harshness as determinants of student performance. [32, 33] In some of the above cases, [30] extended versions of the standard Rasch model – the Multi-Faceted Rasch Model or Polytomous Rasch Model, have been employed to accommodate further factors (the "facets") or response outcomes, but in the absence of any doubt about the soundness of the basic Rasch model.

Most of the above applications of Rasch models rely on their reputation as tools for assessing the objectivity of individual assessments, with lack of closeness of fit of actual item response data to corresponding model data being viewed as an indicator of the need to review exam content or administration. While the role of Rasch models in preserving objectivity has been carried over comparatively less often to the immediate context of standard setting, we note here one such case involving use of the modified Angoff approach so as to provide a more comprehensive argument in support of OSS.

A SEEMINGLY MORE OBJECTIVE APPROACH TO THE MODIFIED ANGOFF PROCEDURE

This case involves the use of hybrid strategies employing the Rasch model alongside the modified Angoff approach. The hybrid strategies exploit the fact that while the Rasch model is primarily intended to serve as an objective model, it can to some extent be manipulated within the standard setting process in such a way that subjective judgements from standard setters under the modified Angoff approach have a considerable role to play. MacCann [27][18] illustrates this attempt to render Angoff standard setting more objective in connection with tests involving dichotomously and polytomously scored items. In the case of dichotomously scored items, which is more common with high stakes assessments in medical education, predictions for individual items of what proportion of MCEs ought to answer the item correctly are combined across judges to form a cut-score for the assessment as with the modified Angoff procedure. The corresponding ability level in logits for the MCE is then determined. This ability level is used as input to the Rasch model in order to obtain the expected proportions corresponding to the likelihood of a MCE passing each of the test items. The expected proportions are then used by judges in a further round of the Angoff process in re-evaluating their own predicted item-specific proportions (PISPs). (Fig. 2)

This integration of Rasch modelling with the Angoff procedure fails to remove the challenge of conceptualizing the MCE and the need for group discussion among judges to allow adjustments to ratings, both of which take place prior to initial input to the Rasch model. Moreover, judges are at liberty to confer further through re-considering their PISPs in the light of Rasch expected item-specific proportions. Here, considerable departure of PISPs from the latter is viewed as undesirable and an indicator that judges have wrongly estimated item difficulties. The practise of calibrating prior PISPs at this stage brings into question their validity; yet, it is these PISPs which have been used to generate the expected item-specific proportions that are now being used as a gold standard to calibrate them (Fig. 2)!

In surveying the cycle of activities (Fig. 2) retrospectively, from the start to the end point, one can see that invalidated PISPs are used to generate rigorous output from the Rasch model so as to ultimately determine new supposedly valid PISPs, which is, of course, non-sensical.

While the Rasch model can serve as a source of data – the expected proportions – relative to which judges may wish to calibrate their predictions, it seems that these proportions are no more valid than the Angoff-based PISPs that are pivotal to their derivation. Correspondingly, it is likely that the relationship between the Angoff-derived input and the implementation of the Rasch model is very much one of the tail wagging the dog.

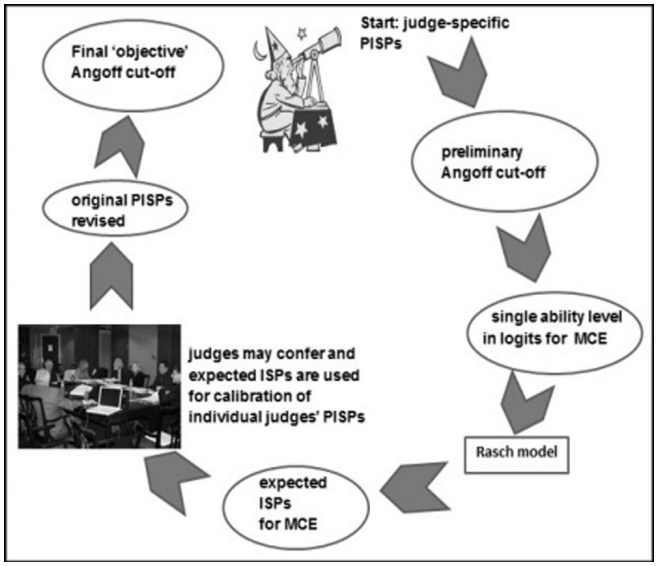

Figure 2: A seemingly more objective Angoff approach. 'MCE, 'ISP' and 'PISP' abbreviate 'minimally competent examinee', 'item-specific proportion' and 'predicted item-specific proportion', respectively.

THE NEED FOR A MORE PURELY OBJECTIVE APPROACH TO STANDARD SETTING

Post-hoc revisions and the threat to validity

The above hybrid approach to the modified Angoff method, like the modified Angoff approach alone, illustrates the more general requirement in standard setting scenarios of judges making successive post-hoc revisions to their decisions. With a wide range of standard setting procedures, including the modified Angoff procedure, there is an inherent need for multiple iterations driven by the results of variance-based measures of judge consistency and experimental item statistics, such as item p values for a group of prior examinees thought to be representative of a borderline. [32, 33][34, 35]. Regrettably, this need undermines the success of the standard setting process in its capacity to produce a *valid* measure of the standard to be set. Furthermore, if particular misconceptions are common among a cohort of judges in meeting the demands of the standard setting task, strategies for reaching consensus, such as the Delphi method and its variants [36] may serve only to reinforce these misconceptions, leading to a false confidence.

Additionally, iterations aimed at inter-judge consistency are open to "social influence effects by dominant committee members" [37] and "the dynamics of the group discussion", including the desire to appear professional. [7] Consequently, convergence of scores can reflect submission to peer pressure rather than identification of a valid standard. There is therefore a lack of convincing evidence to refute the idea that the convergence observed is more of an irreconcilably arbitrary nature than in terms of test content. While it may seem intuitive that accuracy increases with consensus, this alleged trend can only be useful if cut-offs are being polarized

towards the correct anchor – namely, the true standard. Alternatively put, Downing's observation, aimed at the allocation of marks, that "even raters who agree perfectly may be completely wrong" [38] has a clear analogue in judge predictions used to set the standard for an assessment.

In terms of the veracity of the test standard, the pursuit of consensus as an end in itself presents a further problem. In particular, there is a tendency to limit the range of judge expertise in terms of specialty and job role to a level below that which ought to be expressed in the standard. This is the case despite the success of any preliminary effort to recruit panel members who, in their varied capacities, collectively represent the medical profession in a holistic way. Judge predictions are normed against one another, possibly at the expense of the more variegated nature of the true standard which might have been gleaned from different judge backgrounds, such as hospital, clinic and general practise settings. Such a practise is inconsistent with Jaeger's original principle [39] that all groups with a legitimate interest in the standard be included. Correspondingly, the construct validity of the standard set is likely to be compromised.

Problems with normative data

Where item *p* values are used normatively by all judges, this typically has the effect of uniformly lowering their individual cut-offs. [7] Based on the reputation of Angoff approaches for being overly stringent relative to other standard setting procedures, this may be viewed as a welcome effect. However, the usefulness of item p values during the standard setting process has already been brought into question in the educational literature by noting that in representing "average performance of past students on an item", they are not intended to serve as indicators of minimal competence. [7] As noted in earlier work, "speculative judge predictions are ultimately normed to actual test-taker performance, rendering these predictions irrelevant". This is despite the considerable effort made at the start of the standard setting process to allow judges to construct their own conceptual notion of MCE in relation to safe clinical practise. [40] Moreover, if judges are drawn towards the concept of *average student* when trying to conceptualize a MCE, the above standard setting process shares a close affinity with norm-referenced approaches, [41] which are typically regarded as inappropriate for high stakes assessments in medical education. [19]

Clearly, iterative approaches to standard setting also incur considerable financial and administrative burdens in terms of training judges and employing them over several days, including reimbursement of health authorities for releasing employees.

When the above problems of homogenization of the standard and the resultant burdens of workload and expenditure are combined with that of the conceptual challenge of predicting probabilities while visualizing the hypothetical MCE, the threat to test validity and the corresponding need for greater objectivity are particularly clear.

Given this background, there is a strong epistemological basis for identifying an approach to standard setting that is more conceptually transparent and foundationally sound, and through which the test standard may be derived more efficiently and economically, while optimizing construct validity. With this in mind, we now consider OSS more fully.

KEY CHARACTERISTICS OF OSS

Critical to OSS is the requirement that judges focus specifically on content in relation to a given level of competency (e.g. *minimally competent*, *advanced or mastery*). OSS, as presented in this paper, refers specifically to the case of itemized written exams. [40] Given a pool of potentially selectable items (PSIs) representative of a given field, judges must decide between two content-based options for each item. For the given item, the content is either i) essential for a candidate of the chosen competency level to have mastered, where the interpretation of 'essential' has first been discussed and agreed on, or ii) "important, but not essential". [40, 41]

The difficulty rating for each PSI will have already been estimated objectively by Rasch analysis from appropriate (or, calibrated) experimental data using the Rasch model. Thus, the mean or median difficulty rating in logits for each judge's "essential item group" can be obtained. [42, 43] Where the desired proficiency level for the essential items in terms of the number answered correctly is other than 50%, these summary measures are adjusted by a constant in logits; [42] otherwise they remain unchanged. The resultant measures, which are representative of individual judge criteria, are then averaged across judges to obtain a cut-point for the assessment that can in turn be converted to a raw score using either the test characteristic function or curve.

OSS, as presented above, involves choosing between dichotomous responses of the type *essential/inessential*, thus aligning the standard setting task with the above findings in psychology research in relation to conceptual transparency. Furthermore, the decision procedure reflects the foundational intention of Nedelsky, Angoff and Ebel expressed earlier in this paper that the standard be grounded on test content. Thus, OSS approaches ought to have higher face validity than alternative standard setting procedures requiring more complex judgements.

Both OSS and the modified Angoff approach require extensive judge training at the outset of the standard setting exercise. In both cases, judges are required to discuss and define a) the MCE and b) content that is important for MCEs to have mastered to meet the test standard, which can amount to approximately four hours of preliminary work. The major difference in training arises in the decisions expert judges are required to make - Angoff predictions of success in the

form of percentages, versus OSS distinctions between test content which is essential and non-essential. The time and cost savings afforded by OSS arise specifically from dispensing with the consensus-driven iterations that arise once the above steps have been completed. This post-hoc iterative process, which takes hours to complete and often requires several rounds, possibly over several days, is deemed neither necessary nor desirable under OSS. This, in turn, removes the temptation for corners to be cut through use of too few judges in addressing the concerns of cost and efficiency to the possible detriment of test quality. Indeed, it has already been noted that, "Creating and maintaining a pool of examiners who are willing to give up their time for extended periods is difficult." [42]

EXPERIMENTAL FINDINGS

Classification accuracy of the OSS pass mark

Using the beta binomial model to estimate candidate true scores for the American National Board Dental Examination Part II, a 500-item MCQ test, OSS has been shown to have a low false positive rate (<0.001) and false negative rate (0.03) in the classification of examinees as having passed or failed. [44] Such results seem promising given the considerable size of the underlying sample of 1252 examinees.

Comparison of OSS with the Modified Angoff Approach

Using data from five judge panels, OSS has already been compared with the modified Angoff method in high-stakes dental certification written assessments. For a previous year's examination, judges used the modified Angoff method on the first day and OSS on the following day, with the same judges being used in each case. [43] Variance in pass marks across judges was found to be consistently low in OSS relative to the Angoff approach. This pointed to the need for further iterations to allow judges to reach consensus in the latter case but not the former, and exposed the greater conceptual clarity of the former approach. Furthermore, pass marks for panels were less stringent with OSS than with the Angoff approach.

Foundationally, OSS is open to variation in scores across judges so as allow for a holistic standard informed by differing perspectives and backgrounds. Such variation should be within acceptable boundaries and backgrounds and less judge heterogeneity would appear appropriate in the case of more specialist graduate certification examinations [45]. Nevertheless, for the above study, the comparatively greater inter-judge variation (of the order of 10 times greater) in the case of the Angoff approach reflected an undesirably high degree of variation for the Angoff standards to be credible. In keeping with convention for the Angoff approach, this left the standards open to further revision through iterations of the standard setting process, with a view to calibrating judges against one another.

The same study also revealed discrepancies between the two approaches in terms of pass rates. Using quarterly data for past administrations of the above exams over a 3-year period, even where difficulty and ability were controlled for, the Angoff method was found to be very unstable over time in terms of pass rates, with corresponding results from OSS remaining stable. [42] This reflects the fact that unlike in the case of OSS, with the Angoff approach, the standard represented, including the underlying knowledge, skills and abilities, is itself varying over time. Such findings are particularly noteworthy in the light of the conclusion within medical education that "it is not defensible that the standards vary from year to year." [46] Furthermore, they suggest that relative to defining a standard such as *minimal competence*, OSS has much higher construct validity than the modified Angoff approach.

RECOMMENDATIONS FOR FUTURE RESEARCH

The need for transparency in reporting of research

Published findings of the evaluation of OSS are almost entirely restricted to the USA. One exception relates to a study carried out at the Faculty of Engineering and the Built Environment, Universiti Kebangsaan, Malaysia. [47] In this case, scores from a sample of 58 students from the Department of Mechanical and Material Engineering on sitting the Project Design final examination were considered, with an interest in improving entry standards for further study at university level. Regrettably, there are limitations in the clarity of the exposition, leading to uncertainties about the procedural validity of this particular application of OSS and how this specific study can be used to evaluate OSS. Similarly, OSS has been used with the American National Board Dental Hygiene Examination (NBDHE), performance on which is used by individual states as a basis for assessing suitability of candidates for practising dental hygiene. For this 350-item MCQ test, covering a range of disciplines within dentistry, the reliability of the scale has in one recent study been reported as highest at the pass mark (0.97). Further, for the sample of 4,528 candidates, the OSS and actual failure rates (2.8% and 2.4%, respectively) were "reasonably close". [48] However, a careful study of this paper will reveal dependency on the pass mark set by OSS to validate the recommended minimum pass mark previously used for the above examination and vice versa. The inherent circularity in this mutual dependency does not genuinely serve the interests of evaluating the appropriateness of OSS as a gold standard for setting the pass mark, as required. Such findings reinforce the need for comparative studies involving OSS, where the study rationale, design and methodologies are completely transparent and hence open to evaluation by experts in standard setting for educational assessment.

Gaps in the literature

There has been a hiatus in recent research literature in terms of experimental studies exploring concerns about the cognitive complexity of the standard setting task presented

by the modified Angoff approach [10]; yet, such concerns have emerged from highly reputable sources [20, 49, 50].

Existing experimental findings from comparing OSS with the modified Angoff procedure contribute in part to accepting the call to pursue research aimed at addressing these concerns. However, in more fully evaluating the utility and veracity of OSS in high-stakes examinations, there is enormous scope for larger scale experimental studies comparing and contrasting psychometric properties and cost- and time-effectiveness across these two standard setting approaches for itemized tests. Such studies ought to encompass recent standard setting practises which combine the Angoff method and less popular Hofstee method (which recommends lower and upper bounds for an acceptable pass mark and for the percentage of students who should fail). [51]

They should also be extended to include qualitative research relating to ease of use and conceptualization of standard setting methodologies as perceived by judge participants. In particular, there is a call for extensive work involving well-designed studies comparing judge experiences of implementing the modified Angoff procedure and OSS, both overall and more specifically, in terms of the respective standard setting tasks.

For completeness, it is also appropriate to acknowledge the recent arrival of a possible contender to OSS, referred to as the *Objective Borderline Method* (OBM). [52] This method assumes that examiners, in the absence of a panel of judges, are proficient in classifying examinee performance as fail, borderline or pass and thus, can provide a provisional pass mark. Using a probabilistic model, the proportions of examinees falling into the above classes are used to determine an official pass mark for the relevant assessment. It is beyond the scope of this paper to challenge the face validity of the above model. Nevertheless, both OSS and OBM appeal to a probabilistic model as a source of objectivity. Thus, for MCQ tests, it would be of value to compare the stability of the corresponding pass marks over time.

The scope of the above work ought to include both Medicine and allied health professions, such as Dentistry and Nursing.

SUMMARY

Given that Angoff approaches continue to play a prominent role in the standard setting of high stakes undergraduate medical examinations, the concerns expressed earlier in this paper regarding conceptual clarity and validity are of considerable relevance in avoiding the misclassification of aspiring clinicians.

OSS serves as a paradigm for a more purely objective style of standard setting for itemized tests than the modified Angoff approach and is particularly notable for its simplicity, efficiency and potential cost-effectiveness. Current experimental evidence favours the use of OSS over and above the modified Angoff approach. In so doing, this evidence points the practitioner back to the original standard setting task proposed by Angoff, which is more defensible on psychological grounds. It might be argued, therefore, that Angoff's famous footnote served as a red heron in the development of a content-orientated approach to standard setting and that indeed, OSS is more true to the picture as a modified Angoff approach which adheres to Nedelsky, Angoff and Ebel's criterion for standard setting of criterion-referenced assessments.

While published findings in the USA report positive results on implementing OSS, there is also a tremendous door of opportunity open for quantitative and qualitative studies to expand the existing evidence base in favour of OSS. These challenges ought to be considered internationally by researchers in medical education, not only with the above benefits in mind but more importantly, with a view to improving the validity of the standard being set.

ACKNOWLEDGEMENTS AND FUNDING

The authors wish to express their appreciation to the Association for the Study of Medical Education (ASME) for funding the research for this work under their Small Grants Scheme. Some of the findings from this research were presented at the ASME Annual Meeting of 2011, which was held in Edinburgh, UK. Thanks are also due to the journal reviewers who provided insightful comments which enhanced the clarity of content and the focus of this paper.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1. Dudas RA, Barone MA. Setting standards to determine core clerkship grades in pediatrics. Acad Pediatr. 2014;14(3):294-300. Epub 2014/04/29.
2. McKinley DW, Norcini JJ. How to set standards on performance-based examinations: AMEE Guide No. 85. Med Teach. 2014;36(2):97-110. Epub 2013/11/22.
3. Konge L, Clementsen P, Larsen KR, Arendrup H, Buchwald C, Ringsted C. Establishing pass/fail criteria for bronchoscopy performance. Respiration. 2012;83(2):140-6. Epub 2011/10/12.
4. De Champlain AF. Standard setting methods in medical education. In: Shanwick T, editor. Understanding Medical Education. London: Wiley Blackwell; 2014. p. 305 - 16.
5. Angoff W. Scales, norms and equivalent scores. In: Thorndike RL, editor. Educational Measurement. Washington DC: American Council on Education; 1971. p. 506 - 600.
6. Senthong V, Chindaprasirt J, Sawanyawisuth K, Aekphachaisawat N, Chaowattanapanit S, Limpawattana P, et al. Group versus modified individual standard-setting on multiple-choice questions with the Angoff method for fourth-year medical students in the internal medicine clerkship. Advances in Medical Education and Practise. 2013;4:195 - 200.
7. Hurtz GM, Auerbach MA. A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgement constraints. Educational and Psychological Measurement. 2003;63:584 - 601.
8. Ben-David MF. AMEE Guide No. 18: Standard setting in student assessment. Medical Teacher. 2000;22(2):120 -30.

9. Cizek GJ, Bunch MB. Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests. Thousand Oaks, California, USA: Sage; 2007.
10. Plake B, Cikek G. Variations on a Theme: The Modified Angoff and Yes/No Standard Setting Methods. In: Cizek G, editor. Setting Performance Standards: Foundations, Methods and Innovations. 2nd ed. New York: Routledge; 2012. p. 181 - 99.
11. NAE. Setting Performance Standards for Student Achievement. Washington DC: National Academy of Education; 1993.
12. Wang N. Use of the Rasch IRT model in standard setting: An item-mapping method. Journal of Educational Measurement. 2003;40(3):231 - 53.
13. Impara J, Blake B. Standard setting: An alternative approach. Journal of Educational Measurement. 1997;34(4):353 - 66.
14. Tversky A, Kahneman D. Judgement under uncertainty: heuristics and biases. Science. 1974;185:1124 - 31.
15. Gupta P, Dewan P, Singh T. Objective Structured Clinical Examination (OSCE) revisited. Indian Pediatr. 2010;47(11):911-20. Epub 2010/12/15.
16. Ebel R. Essentials of Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall, Inc; 1979.
17. Nedelsky L. Absolute grading standards for objective tests. Educational and Psychological Measurement. 1954;14(1):3 - 19.
18. Cizek G. Reactions to National Academy of Education Report, Setting Performance Standards for Student Achievement. Washington, DC: 1993.
19. Examinee-centred vs. task-centred standard setting. In: Kane M, editor. Joint Conference on Standard Setting for Large-Scale Assessments; Washington, DC: National Assessment Government Board and Center for Education Studies; 1995.
20. McManus I, Chis L, Fox R, Waller D, Tang P. Implementing statistical equating for MRCP(UK) parts 1 and 2. BMC Medical Education. 2014;14(204).
21. Tavakol M, Doody GA. Making students' marks fair: standard setting, assessment items and post hoc item analysis. International Journal of Medical Education. 2015;6:38 - 9.
22. Rasch G. Probabilistic Models for Some Intelligence and Some Attainment Tests, Expanded Edition. Chicago: University of Chicago Press; 1960.
23. Remo O, Nering M. Polytomous Item Response Theory Models. Liao T, editor. Thousand Oaks, London, New Delhi: SAGE Publications; 2005.
24. Linacre JM. Sample size and item calibration [or person measure] stability. Rasch Measurement Transactions. 1994;7(4):328.
25. Yen WM, Fitzpatrick AR. Item Response Theory. In: Brennan RL, editor. Educational measurement. New York: American Council on Education and MacMillan Publishing Company; 2006. p. 111 - 53.
26. Stone GE. Introduction to the Rasch Family of Standard Setting Methods. In: Smith EVJ, Stone GE, editors. Criterion Referenced Testing: Practise Analysis to Score Reporting Using Rasch Measurement Models. Maple Grove, Minnesota, USA: JAM Press; 2009a. p. 138 - 47.
27. MacCann RG. Standard setting with constructed and dichotomous response items: some Rasch model approaches. In: Smith EVJ, Stone GE, editors. Criterion Referenced Testing: Practise Analysis to Score Reporting Using Rasch Measurement Models. Maple Grove, Minnesota, USA: JAM Press; 2009. p. 251 - 75.
28. Grosse ME, Wright BD. Setting, evaluating and maintaining certification standards with the Rasch mode. Evaluation & The Health Professions. 1986;9(3):267 - 85.
29. O'Mara DA, Canny BJ, Rothnie IP, Wilson IG, Barnard J, Davies L. The Australian Medical Schools Assessment Collaboration: benchmarking the preclinical performance of medical students. Med J Aust. 2015;202(2):95-8. Epub 2015/01/30.
30. Sebok SS, Roy M, Klinger DA, De Champlain AF. Examiners and content and site: Oh my! A national organization's investigation of score variation in large-scale performance assessments. Adv Health Sci Educ Theory Pract. 2015;20(3):581-94. Epub 2014/08/29.
31. Tor E, Steketee C. Rasch analysis on OSCE data : An illustrative example. Australas Med J. 2011;4(6):339-45. Epub 2011/01/01.
32. Peeters MJ, Sahloff EG, Stone GE. A standardized rubric to evaluate student presentations. Am J Pharm Educ. 2010;74(9):171. Epub 2011/02/09.
33. Till H, Ker J, Myford C, Stirling K, Mires G. Constructing and evaluating a validity argument for the final-year ward simulation exercise. Adv Health Sci Educ Theory Pract. 2015. Epub 2015/03/27.
34. Norcini J. Equivalent pass/fail decisions. Journal of Educational Measurement. 1990;27(1):59 - 66.
35. Norcini JJ, Shea JA, Kanya T. The effect of various factors on standard setting. Journal of Educational Measurement. 1988;25(1):57 - 65.
36. Jaeger RM. An iterative judgement process for establishing standards on competency tests: theory and application. Educational Evaluation and Policy Analysis. 1982;4(4):461 - 75.
37. Plake B. Factors influencing intrajudge consistency during standard setting. Educational Measurement: Issues and Practise. 1991;10(2):15 - 6, 22, 5-6.
38. Downing S. Reliability: on the reproducibility of assessment data. Medical Education. 2004;38:1006 - 12.
39. Zeiky MJ. So much has changed. How the setting of subscores has evolved since the 1980s. In: Cizek GJ, editor. Setting performance standards: concepts, methods and persepectives. Mahwah, N J: Lawrence Erlbaum Associates; 2001. p. 19 - 52.
40. Stone GE, Koskey KLK, Sondergeld TA. Comparing construct definition in Angoff and objective standard setting models: playing in a house of cards without a deck. Educational and Psychological Measurement. 2011;71(6): 942-962.
41. Mortaz Hejri S, Jalili M. Standard setting in medical education: fundamental concepts and emerging challenges. Med J Islam Repub Iran. 2014;28:34. Epub 2014/09/25.
42. Stone GE. Objective standard setting (or truth in advertising). Journal of Applied Measurement. 2001;2(2):187 - 201.
43. Stone GE, Beltyukova S, Fox CM. Objective standard setting for judge-mediated examinations. International Journal of Testing. 2008;8:180-96.
44. Tsai T-H, Neumann LM, Littlefield JH. Validating the standard for the National Board Dental Examination Part II. Journal of Dental Education. 2012;76(5):540-4.
45. Cizek GJ. Standard-setting guidelines. Educational Measurement: Issues and Practise. 1996;15(1):13 - 21.
46. Schuwith L, van der Vleuten C. How to Desigh a Useful Test: the Principles of Assessment. Edinburgh: Association for the Study of Medical Education; 2006.
47. Khatimin N, Abd Aziz A, Zaharim A, Sahari J, Rahmat RAOK, IEEE. Setting the Standard for Project Design course using Rasch Measurement Model. 2013 IEEE Global Engineering Education Conference2013. p. 1062-5.
48. Tsai T-H, Dixon BL. Setting and validating the pass/fail score for the NBDHE. Journal of dental hygiene : JDH / American Dental Hygienists' Association. 2013;87(2):90-4.
49. Pelligrino J, Jones L, Mitchell K, editors. Grading the Nation's Report Card. Washington, DC: National Academy Press; 1999.
50. Shepard L, editor. Implication for standard setting of the National Academy of Education evaluation of National Assessment of Educaional Progress achievement levels. 1995. Washington, DC: US Governmen Printing Office; 1995.
51. Ali K, Coombes L, Kay E, Tredwin C, Jones G, Ricketts C, et al. Progress testing in undergraduate dental education: the Peninsula experience and future opportunities. Eur J Dent Educ. 2015. Epub 2015/04/16.
52. Shulruf B, Turner R, Poole P, Wilkinson T. The Objective Borderline method (OBM): a probability-based model for setting up an objective pass/fail cut-off score in medical programme assessments. Advances in Health Sciences Education. 2013;18(2):231-44.

Source of Support: Nil, Confl ict of Interest: None declared