BMC Medical Education

RESEARCH ARTICLE Open Access

Insights into the Angoff method: results from a simulation study

Boaz Shulruf[1*], Tim Wilkinson[3], Jennifer Weller[2], Philip Jones[1] and Phillippa Poole[2]

Abstract

Background: In standard setting techniques involving panels of judges, the attributes of judges may affect the cut-scores. This simulation study modelled the effect of the number of judges and test items, as well as the impact of judges' *attributes* such as accuracy, stringency and influence on others on the *precision* of the cut-scores.

Methods: Forty nine combinations of Angoff panels ($N = 5$, 10, 15, 20, 30, 50, and 80) and test items ($n = 5$, 10, 15, 20, 30, 50, and 80) were simulated. Each combination was simulated 100 times (in total 4,900 simulations). The simulation was of judges attributes: stringency, accuracy and leadership. Impact of judges attributes, number of judges, number of test items and Angoff's second (compared to the first) round on the precision of a panel's cut-score was measured by the deviation of the panel's cut-score from the cut-score's true value.

Results: Findings from 4900 simulated panels supported Angoff being both reliable and valid. Unless the number of test items is small, panels of around 15 judges with mixed levels of expertise provide the most precise estimates. Furthermore, if test data were not presented, a second round of decision-making, as used in the modified Angoff, adds little to precision. A panel which has only experts or only non-experts yields a cut-score which is less precise than a cut-score yielded by a mixed-expertise panel, suggesting that optimal composition of an Angoff panel should include a range of judges with diverse expertise and stringency.

Conclusions: Simulations aim to improve our understanding of the models assessed but they do not describe natural phenomena as they do not use observed data. While the simulations undertaken in this study help clarify how to set cut-scores defensibly, it is essential to confirm these theories in practice.

Background

Standard setting is an important aspect of assessment, with the literature describing a plethora of methods. Although each has unique features, most standard setting methods use panels of expert judges to determine the cut-scores between the different performance categories [4, 5, 14, 47, 50]. Among the judge-based standard setting methods, Angoff's method (henceforth Angoff) and variants, have been used in a range of educational settings [2, 4, 13, 59]. Commonly, Angoff is a process used to estimate performance standards at the pass-fail level; i.e. a process aiming to 'separate the competent from the non-competent candidate' ([5], p. 120). In this process, each judge estimates the proportion of minimally competent examinees who would give a correct answer to each of the items. Those estimates are then summed across items for each judge, with the average of the sums across judges determining the test cut-score [2]. A variant, the modified Angoff, includes a second round of judgements after the judges have seen their peers' judgements. This has been shown to increase inter-judge agreement [13]. Furthermore, [15] demonstrated that Angoff group discussion, which did not include test results, decreased the variance of within-panel estimation of the proportion of correct responses per item; however these discussions did not decrease the differences between the judges' estimates and the observed proportion of correct values.

Research on the utility of Angoff suggests that the cut-scores generated by a panel are affected by the panel's composition, particularly the number of judges and their levels of expertise [7, 17, 32, 66, 70]. Numerous modifications have been introduced to the original Angoff method in order to improve the defensibility of the resulting cut-scores [7, 21, 37, 38, 54, 55]. These

* Correspondence: b.shulruf@unsw.edu.au
[1] University of New South Wales, Sydney, Australia
Full list of author information is available at the end of the article

modifications include providing additional information to the judges, such as judgements made by peers, normative examinee data or pre-judgement training [13, 54]. However, there remains uncertainty about the relative impacts of the number of test items and judges, or judges' attributes, on cut-scores. The recommended number of judges for Angoff ranges from 5 to 30 [17, 26, 32, 33, 44, 48]. Nonetheless, [36] demonstrated that if judges were randomly sampled from a large pool of qualified judges, at least 87 judges were needed, with 95% probability, to ensure that the cut-score estimation error did not exceed one test item. The impact of the number of test items on the cut-score appeared to be small [26, 33] even when subsets of items taken from the same tests were considered [25]. The impact of the number of items on the validity of the cut-score determined by Angoff panels has not been widely studied. A common view is that the resulting cut-score will be more accurate as the subject expertise of the judges increases; nonetheless, that assertion has not been confirmed empirically [10, 36, 64, 66, 70].

A major challenge in the literature on standard setting is that there is no 'gold standard' for standard setting [71]. Furthermore, the Consensus Statement and recommendations from the Ottawa 2010 Conference suggest that validating an assessment by comparing one assessment criterion with another has 'lost ground,' since the other assessment criterion also needs validation. This is, in effect, an endless or perpetual process [56]. With the exception of a rough estimate of expertise, i.e. experts vs. non-experts, most evidence on the quality of standard setting is extracted from data from different tests and using different panels where judges' attributes were not measured.

The current study aims to explore the potential effect of panel constitution and expertise on standard setting. In particular, this study models the impact of the number of judges, the number of items and judges' *attributes* on the *precision* of the resulting cut-scores as well as the impact of a second round of Angoff on the precision of the cut-score. The research questions are:

1. Is there an optimal number of judges and items for the Angoff standard setting process?
2. What is the impact of a judge's *attributes* on the *precision* of the cut-score?
3. To what extent does the second round of decision-making (where judges' decisions are affected by the composition and decisions of other panellists, but not by test parameters) improve the *precision* of the cut-score?

Methods

To address these questions, this study used simulated data. By their nature, simulated data, a priori, establish the correct (true) value for the cut-score, and hence provide accurate and valid criterion validity [56] for assessing a model [22] of a standard setting method, in this case the Angoff method. For clarity, two cut-scores are discussed in this manuscript: (1) the 'true' cut score which is determined by the simulation as described below; and (2) the 'cut-score' which is yielded from the simulated judges' decisions.

By using simulated data, it is possible to compare the cut-scores determined by Angoff panels of simulated judges with the 'true' cut-score as set by the simulation parameters. Having a 'true' cut-score means that two fundamental assumptions must underlie this study: (1) there is a cut-score that distinguishes competence from incompetence [5], for example the common definition 'minimally competent examinee' ([73], p. 219); (2) an examinee must be either competent or incompetent, but cannot be both or neither.

This study simulates judges' attributes and the impact of these attributes on the cut-score yielded from the Angoff method. To simulate the effect of *attributes* on the *precision* of the cut-score required the generation of a judge's cut-score for each item. This was made under some assumptions:

(a) A judge's expertise was assumed to be positively associated with greater accuracy (i.e. the smaller the deviation of the judge's cut-score from the true cut-score, the more expert the judges were); thus 'Accuracy' in this study is equivalent to expertise;
(b) Based on previous evidence [64, 66, 70] experts are regarded as more stringent than novices;
(c) Experts are more likely to have greater influence within the panel [10], which is designated as 'Leadership';
(d) A judge's estimation of the cut-score is affected by a combination of their personal attributes and a random error [53]. Note that random errors are independent, normally distributed around the true value, with their sum equal or very close to zero [31];
(e) A judge's *attributes* are independent of item difficulty [24, 27, 31, 52, 60]. As level of expertise and content-specific knowledge impact on judges' decisions [66], level of expertise was included in the analysis;
(f) There is no predefined way of determining the relative impact of stringency and accuracy on a judge's decisions. This assumption was made since no evidence was found in the literature suggesting otherwise;
(g) 'Leadership' (influence of one judge on others in the second round) is associated with two independent components: the first is a general social attribute of leadership which is independent of expertise; the second is related to expertise, since judges are likely to change their views based on information deemed to be correct, as do experts [10, 16].

Table 1 Simulation parameters for generating a score made by Judge j for item i

Notation	Parameter	Random function
A	Judge's Accuracy	Normal (μ = 0; σ = 1)
S	Judge's Stringency	Normal (μ = 0; σ = 1)
L	Judge's Leadership	Normal (μ = 0; σ = 1)
Wa	Accuracy's Weight	(Normal (μ = 0; σ = 1) + 3)/3
Ws	Stringency's Weight	(Normal (μ = 0; σ = 1) + 3)/3
Ji	Judge's raw score of item i	Normal (μ = 0; σ = A_j)
Sji	Judge's Stringency for item i	Normal (μ = S_j; σ = 1)
Sa	The impacts of Accuracy on Stringency[a]	Normal (μ = 1-A_j; σ = 1)
La	The impacts of Accuracy on Leadership[a]	Normal (μ = 1-A_j; σ = 1)
Lji	Judge's Leadership in the second round	Normal (μ = L_j; σ = 1)

[a] Impact of accuracy is the component of Accuracy (expertise) that contributes to Stringency or Leadership

The *attributes* that were simulated in this study included:

1. Accuracy: the accuracy of a judge's cut-score was simulated by varying the size of the random error component in the judge's cut-score [23].
2. Stringency: the extent to which a judge's cut-score is affected by bias by being more stringent or lenient was simulated by adding (subtracting) a systematic error to all judges' decisions where the size of the error determines the level of stringency/leniency [23].
3. Leadership: The extent to which a judge's cut-score in the second Angoff round is influenced by the cut-scores determined by other judges was simulated by determining the relative contribution (using weighted average) of each judge to the panel's overall decision. This attribute could be independent of, or associated with, level of expertise. These associations are discussed further below.
4. The relative impact of Stringency and Accuracy on a judge's decisions: the extent to which each of these attributes (Stringency and Accuracy) affects or dominates the judge's decision on the cut-score was simulated by the proportional impact of each attribute. These simulation parameters were independent and were included in the simulation to allow different impacts of Stringency and Accuracy on judges' decision.

The simulation applied standardised measures whereby the correct (true) cut-score for each item (and consequently for the whole test) was set to zero. Hence this study measured only the judges' decisions and not the difficulty of the items.

The data simulation and data analysis were undertaken using SPSS V22. Data generation parameters used in this simulation are described in Table 1. This table summarises how a judge's *attributes* were simulated.

Formulae for calculating the judges' mean score for item$_i$, which always has a *true* value = 0, are described below. Score1 is the cut-score generated by the panel in round 1 and Score2 is the cut-score generated by the panel in round 2 when only impact of other judges i.e. 'Leadership' was added to Score1. Other possible impacts (e.g. students' actual results) used in different versions of Modified Angoff [37, 49] were not simulated in this study.

It was necessary to simulate weights for Leadership since the second round differed from the first round only by Leadership. 'Leadership' was manifested by the weight given to a particular judge's decision when averaging the panel's cut-score.

The *attributes* described above were generated for each judge within each panel and used to generate that judge's scores for each item. Note that these determinants were constant (although derived from a random number function) across all scores for all test items given by the same judge as they described judge's *attributes* only, but not the scores given by the judge, which was a random number generated by those determinants.

$$\mathrm{Score1} = \left[\textstyle\sum_{(j=1\ \mathrm{to}\ n;\ i=1\ \mathrm{to}\ k)}\left[\left(\mathrm{Ji}_{ji} * \mathrm{Wa}_{ji} + \mu\left(\mathrm{S}_{ji},\ \mathrm{Sa}_{ji}\right) * \mathrm{Ws}_{ji}\right)\right.\right.$$

$$\left.\left./\textstyle\sum\left(\mathrm{Wa}_{ji},\ \mathrm{Ws}_{ji}\right)\right]\right]/(n * k)$$

$$\text{Overall Leadership of judge}_j = \mu\left(\left(\mathrm{L}_{ji},\ \mathrm{La}_{ji}\right)\ + 3\right)/3$$

$$\mathrm{Score2} = \left[\textstyle\sum_{(j=1\ \mathrm{to}\ n;\ i=1\ \mathrm{to}\ k)}\left(\mathrm{Score1}_{ji} * \left[\mu\left(\left(\mathrm{L}_{ji},\ \mathrm{La}_{ji}\right)\ + 3\right)/3\right)\right]\right]/(n * k)$$

n = no. of judges; k = no. of items

All simulation parameters were standardised to have standard deviation of 1 and mean of 0 or 1 as appropriately required. The results generated by the simulations present the effect of the judges' *attributes* on the precision (deviation from the true cut-score) of the resulting cut-scores, which was set to zero in all simulations.

The Angoff simulations were performed for a range of judges (5, 10, 15, 20, 30, 50, and 80) and a range of items (5, 10, 15, 20, 30, 50, and 80). These choices were based on previous research showing that the addition of items or judges has greater impact when the numbers are small and the association is not linear [26, 33]. For each simulation, a new panel was generated where each judge had a unique set of *attributes*.

Overall there were 49 simulation combinations, each comprising 100 simulations.

Statistical analysis

The means and 95 % CI of the means of the cut-scores derived from the 49 sets of 100 panels were calculated and compared. A paired t-test was used to measure the difference in Angoff cut-score precision between the two rounds. Graphical presentations were used to demonstrate the interactions between judges' *attributes* and *precision*.

Ethical statement

This study used simulated data only thus no ethical approval was required.

Results

The results were derived from 4,900 simulated Angoff panels comprising 100 simulations of each of the 49 combinations of the number of judges and of items.

(1) *Is there an optimal number of judges and items for the Angoff standard setting process*?

The results (Fig. 1) show that the 95 % CI of the mean cut-scores for each of the 100 panels in all judge-item combinations included the true score (zero). There was no relationship observed between the mean cut-score and the number of judges or items. This lack of relationship was evident in the cut-scores both from round one and round two.

However, increasing the number of judges was associated with narrower confidence intervals, irrespective of the number of items.

Figure 2 shows the effect of increasing the number of judges on the *precision* of the panels' cut-scores for tests with different number of items. For tests with 10 or fewer items, increasing the number of judges significantly improves the *precision*, although there is not more to be gained when the number of judges exceeds 30. For larger tests, increasing the number of judges beyond 15–20 has little effect on improving *precision*.

Note that in these analyses (Figs. 1 and 2), judges' *attributes* are not considered and they had no impact since, based on the simulation parameters, their overall impact on any set of 100 panels is equal or very close to zero.

(2) *What is the impact of judges' attributes on the precision of the Angoff cut-score*?

The first attempt to answer this question was made by measuring the partial correlation (controlled for number of judges and number of items) between the

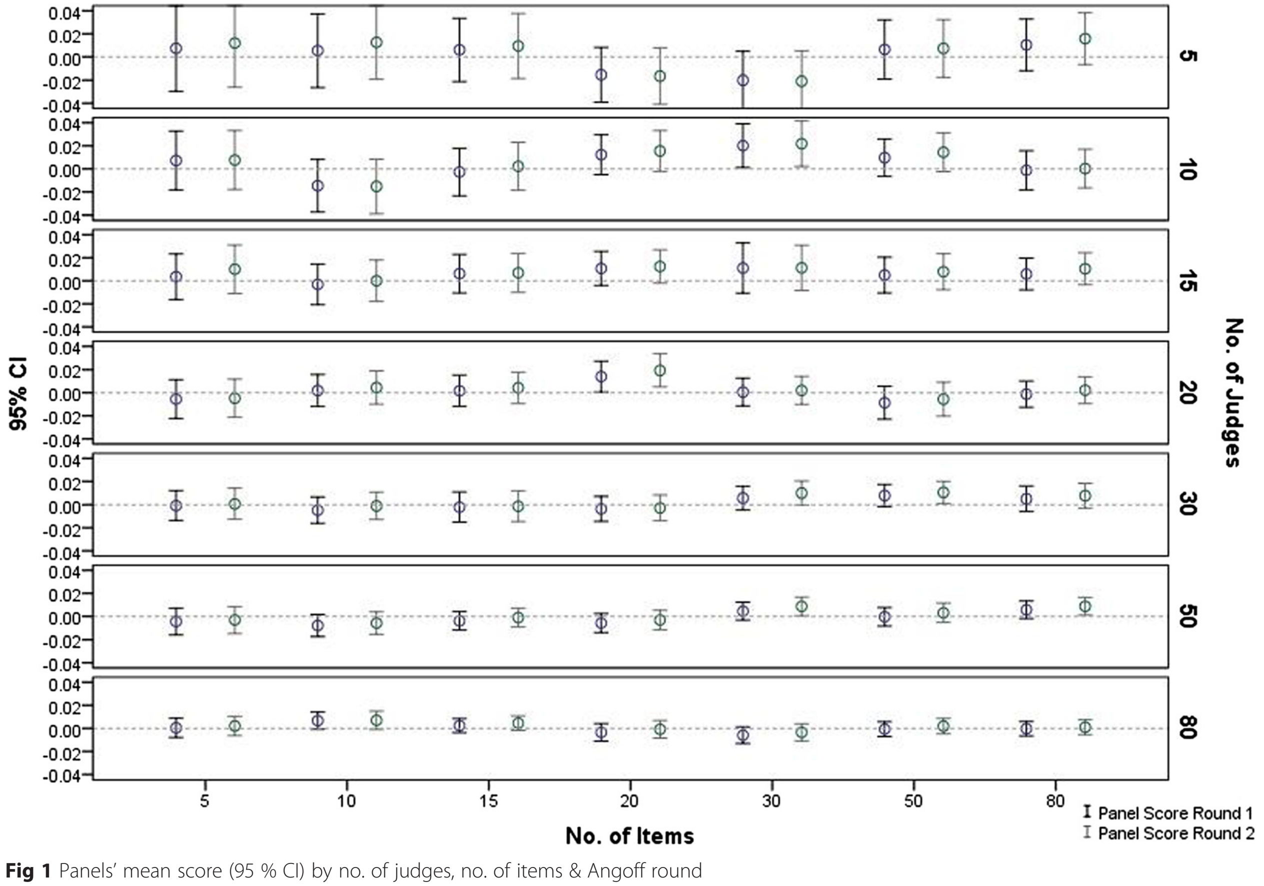

Fig 1 Panels' mean score (95 % CI) by no. of judges, no. of items & Angoff round

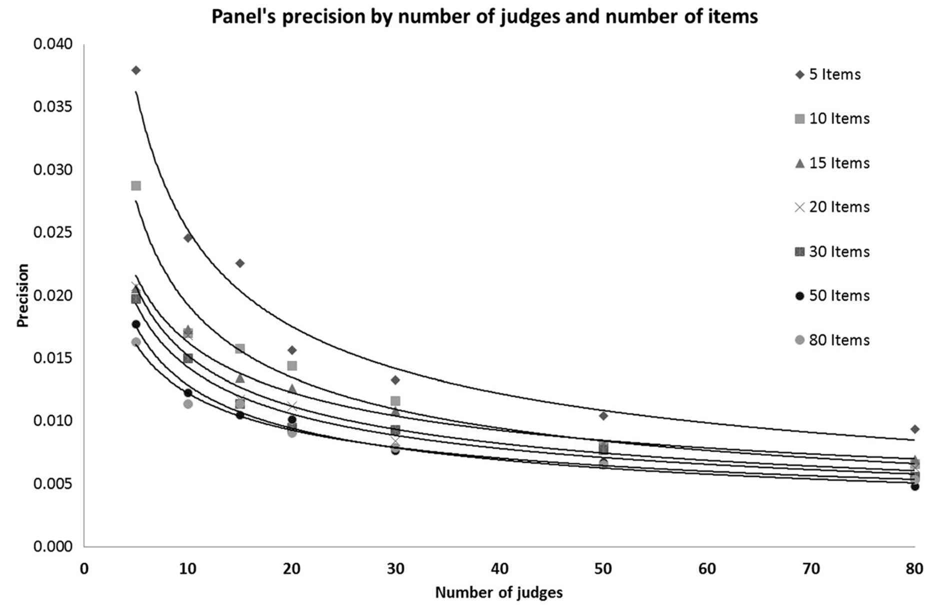

Fig 2 Panels' precision by no. of judges and no. of items

means of the panel's cut-scores and the means of within-panel judges' *attributes* (Accuracy, Stringency and Leadership). The correlations were negligible and statistically insignificant. The panels were then classified into deciles (each consisting of 490 panels) based on judges' *attributes*, to allow identification of any non-linear association. Based on the simulation parameters, the impact of the number of judges and number of items within each decile is very close to zero, thus there was no need to control for those variables.

The results as shown in Fig. 3 indicate that the association between judges' *attributes* and the *precision* of the cut-score is complex and non-linear. Figure 3 demonstrates that panels' cut-scores are most precise (i.e. closer to the true value of zero) when the judges are neither too stringent nor too lenient. However, the interaction of judges' Accuracy and Stringency showed that when the panel is neither too stringent nor too lenient, the impact of panels' mean Accuracy (i.e. the within panel agreement) on the cut-score *precision* is small.

(3) *To what extent does the second round of decision-making improve the precision of the Angoff cut-score*?

A paired t-test was employed to measure the difference between absolute cut-scores from the two rounds. The difference was statistically significant but practically negligible (Round 1 = .060; Round 2 = .0613; $N = 4900$; $p < .001$; Cohen's d = −0.083). Additional analysis measured the correlation between judges' level of agreement (SD within panel) and panel precision (cut-score in absolute values). Partial correlation (controlling for number of judges and number of items) between within-panel SD and panel precision (cut-score absolute value) was used to measure the impact of judges' agreement on the panel cut-score *precision*. The correlation was statistically significant but low ($r = .226$, $p < .0001$) indicating that, although there was a correlation, the within-panel SD explained only 5.1 % of the variance in the cut-score *precision*.

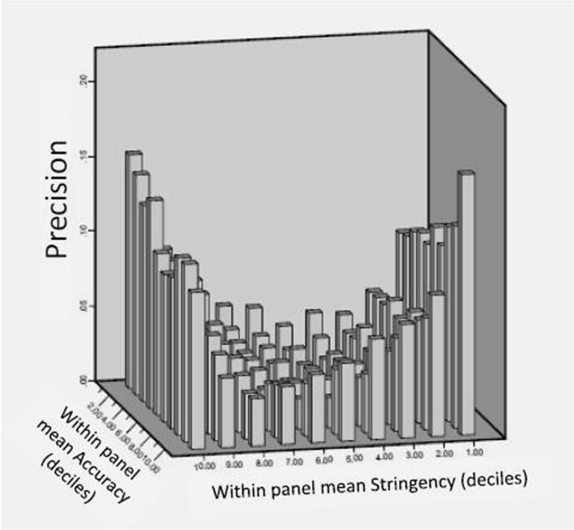

Fig 3 Panels cut-score accuracy by within-panel judges' Accuracy and Stringency

Discussion

This study used simulated data based on 4900 unique panels of judges, which allowed measurement of the difference between the panels' determined Angoff cut-scores and the 'true' cut-score. The main findings were:

(a) Increasing the number of judges reduces the variation in the panel's cut-scores but, more importantly, also increases the *precision* of the panel's cut-score; however, the effect on *precision* was less evident for tests with a large number of items;
(b) Judges' Stringency and, to a lesser effect, judges' Accuracy affect the cut-score *precision*; and
(c) Applying the second round of Angoff process without consideration of examinees or test data does not have a meaningful impact on cut-score *precision*.

The findings are discussed in three sections. The first discusses the merit and the appropriateness of the simulation; the second discusses the findings and their implications for researchers and practitioners; and the third section discusses the merit and limitations of this study and possible directions for further research.

The simulation

Simulated data have been used previously in educational assessment research for knowledge-based tests [18, 34, 52, 60, 62] and for performance-based assessment [40]. However, simulation studies in the field of standard setting are scarce and none was found that simulated judges' decisions based on their simulated attributes and comparing them with a simulated 'true value' [7, 34, 45, 72]. Most previous simulation studies in this field simulated student performance/ examination scores to be used by Angoff panels comprising real judges, yet none of these studies measured judges' attributes and their impact on the cut-score precision [7, 74]. B. Clauser et al. [15] compared the judges' estimates of proportion correct answers with empirical data of examinees' proportion correct answers. This approach, although important, measures the judges' ability to estimate examinees' performance on a particular test, but without any empirical evidence to suggest the cut-score that distinguished competence from incompetence [59]. The current study builds upon previous works [3, 15, 33, 43, 61] and extends the use of simulation in this field by simulating judges' *attributes* that are assumed to affect their decisions, as well as measuring the *precision* of the cut-score by comparing the panels' determined cut-score with the 'true' cut-score.

All previous studies identified in the literature used the variance within judges (or agreement among) as a measure of accuracy or precision. Using such a measure means that if a panel of judges was very stringent but all agreed with each other their agreed cut-score would be deemed more accurate than a cut score yielded by a balanced panel comprising some stringent and some lenient judges, which naturally would yield a larger variance. In real life there is no way to know the true cut-score that distinguishes between competence and incompetence, hence standard setting is employed. For example ([11], p. 158) presented data showing that three different panels estimating the same items yield different agreed cut-scores and different inter-rater variance even when using the same standard setting method (Angoff or Nedelsky). Other studies, (e.g. [17, 33, 37, 48, 65]) which used generalizability analysis to measure the replicability of an Angoff procedure, concluded that a large portion of the overall error variance came from the judges, yet they had no gold standard with which to measure deviation from the true cut-score. This is obvious since generalisability analysis is based on sources of errors while assuming that the mean is very close to the true score [9]. When measuring the precision of a standard setting process, simulation studies like the one presented in this paper, have the unique advantage of including the true cut-score as a valid standard for comparison [58].

The rationale justifying the simulation of each of the variables is discussed in detail in the Method section and not repeated here. However, is it valid to simulate judges' attributes? Verheggen et al. [64] demonstrated that in standard-setting, a judge's individual decision on an individual item reflect the 'inherent stringency of the judge and his/her subject-related knowledge' ([64], p. 209). This notion was widely mentioned in the literature [17, 66, 70]. Thus, in measurement terms [30], if all items are equally difficult (i.e. difficulty level =0) then the resulting cut-score is comprised of the sum of biases i.e. Judges' Stringency and sum of random errors i.e. Accuracy and other random errors. Since previous studies suggest that experts are more stringent than non-experts, [64, 66, 70] and are deemed to have greater influence within the panel [10], we included these assumptions in the simulation parameters. The absolute extent to which each of the attributes affects the judgement is unknown, thus the simulation was comprised of standardised parameters ($SD \cong 1$) to allow the relative impacts of each parameter on the cut-scores to be ascertained. Note that like all simulation studies, the current study measures interactions for given simulated conditions, for better understanding of an assessment model. This study is not about measuring nature [22]. However, this study is similar to research using real data, in that one study measures impact observed on a particular sample and a different study applies similar measures on a different sample. Often the results are different, yet the difference does not suggest that one study is more correct than the other. Given the concordance with previous studies that used real data [33], it is suggested that the results of this simulation study would be applicable to any population of judges with attributes not unlike what was simulated in this study.

Overall, a simulation study always yields results which are determined by the simulation parameters. The contribution of this study to the standard setting literature is that it measures the impact of judges' *attributes* at the individual level on the *precision* of the panel's cut-score. To our knowledge, these associations have never been measured before, either by using simulated or observed data. The concordance of the results of this study with previous studies, particularly where results could be compared (e.g. Fig. 2 vs. work of Hurtz and Hertz [33], Fig. 1), support the validity of the simulation assumptions and parameters, thus adding strength to the study findings.

Implications of the results

Angoff is often used to set standards in large scale educational assessments [8, 59]. Within the context of medical education, Angoff has been applied to tests of medical knowledge (e.g. MCQ's) [28, 49], or clinical skills examinations (e.g. OSCE) [20, 37, 55].

In clinical examinations (e.g. OSCE), the number of items (or stations) may be between 10 and 20 [6]. Thus, given that increasing the number of items is unlikely, for reasons of feasibility, our results suggest that if Angoff were used, an optimum combination would be about 30 judges for 10 items, with a minimum of 20 judges for 15 items or more. For MCQs, where the number of items is large [69], a minimum of 15 judges should suffice for setting up a defensible Angoff cut-score for examinations consisting of 80 items or more (Fig. 2). It is noted that increasing the number of items provided more data points , thus higher reliability [1] and therefore also is likely to increase precision.

These findings are within the range recommended in the literature, suggesting that an acceptable cut-score could be reached if 5–25 judges were employed [17, 33, 41, 46, 48]. Since there is no gold standard for any definition of 'what is good enough' in standard setting [19, 76], applying Angoff with different numbers of judges might be justifiable depending on the context of the examinations.

Previous studies using observed data have determined Angoff precision by the variance across the judges [37, 48]. Other studies that used observed data used IRT parameters or cut-scores generated by alternative methods to estimate the quality of the Angoff generated cut-scores [63, 75]. These methods are appropriate when observed data are used. In the current study, *precision* was determined by the deviation of the panel's cut-score from the 'true' cut-score. The difference between these definitions is more than semantic. Jalili et al. [37] and others [17, 51] used indirect measures to estimate validity as for example, Jalili et al. [37] stated 'We do not have a reference standard by which to test validity'. Their elegant solution was to use correlation between the panels' cut-scores and mean observed scores (scores given to examinees by the examiners) for each item as a measure for estimating validity. The current study has the advantage of having a reference standard by which to test validity since it was included in the simulation parameters (true cut-score = 0). Our finding that the correlation was low ($r = .226$, $p < .0001$) indicates that although there was a correlation, the within-panel SD (judges agreement) explained only 5.1 % of the variance in the cut-score *precision*. This finding is important as it suggests that although identifying the source of error (i.e. in generalizability studies) is a valid way to measure the reliability of a standard setting method [39], using the true cut-score, or an acceptable proxy of it (if real data are used), is an invaluable reference for measuring validity [57]. Consequently, this finding supports a re-thinking of the composition of Angoff panels.

The literature suggests that the Angoff judges should be experts [32], yet it recognises that experts are more stringent and may have greater influence on other judges [10, 64, 66, 70]. Fig. 3 provides some insight into this discrepancy by demonstrating the interaction between Stringency and Accuracy (being an expert). It seems that panels that are neither too stringent nor too lenient are more accurate as they are less prone to bias. However, the level of Accuracy (individual's ability to estimate the correct cut-score) has only small impact on the panel's cut-score *precision*. This is plausible, since the cut-score is determined by the mean of all judges' scores [30]. Without bias in the judgement (assuming Stringency is held constant), the mean score achieved by the judges gets closer to the true value as the number of judges increases [30]. The impact of Stringency on *precision* is obvious (as it was one of the simulation parameters) but it also suggests that a panel which has only experts or only non-experts would yield a cut-score which is less *precise* than a cut-score yielded by a mixed-expertise panel (Fig. 3), particularly given the already-documented association between stringency and expertise [10]. Overall these findings suggest that optimal composition of an Angoff panel should include a diverse range of judges in terms of expertise and stringency (if known). Given the small impact of judge agreement on cut-score *precision* (variance explained = 5.1 %), this practice is recommended despite the likelihood of increasing within-panel judges disagreement.

This study found that the impact of a second Angoff round, where judges may be influenced by others (i.e. influence of 'Leadership'), is negligible. Although this finding was negligible even when measured by standardised effect size (Cohen's d = −0.083) it needs to be interpreted with caution particularly since the measures are all standardised and the second round was different from the first *only by the influence of judges*. This finding is supported by previous empirical studies demonstrating minor differences between two Angoff rounds [15, 61]. Other factors, such as presentation of test data, were not included in this study. It is possible that a different weighting method would have yielded a larger impact and this should be tested in future studies. The literature justifies the second round as a way to increase agreement among the judges [28, 32, 37, 43], yet as indicated above, increasing the within-judges agreement may have little impact on cut-score *precision*, which explains the observed lack of impact of a second round on the cut-score *precision*. The inevitable conclusion from these somewhat surprising results suggests that, provided there are enough judges, the original unmodified Angoff's method [2] is robust enough and the discussion among the panellists does not significantly improve the precision of Angoff's cut-score.

Nonetheless, the modified Angoff methods that provide additional information on the test performance itself (e.g. item and student parameters based on IRT analyses) [15, 42, 43, 63, 68] are welcomed. Such modifications are likely to increase judges' precision without impact on Stringency, as this additional knowledge is related to test parameters only and not to level of expertise.

Study limitations

This study has limitations, the main one being that it is a simulation study. The validity of the findings depends on the validity of the data simulation, especially the variables and the assumptions. We assumed that the judges' *attributes* are normally distributed, rather than non-parametric. Naturally, it is possible that a particular examination and/or particular set of examinees and/or particular set of judges in real life would have different attributes from what is described in this study and thus the recommendations of this study would not be applicable for them. However, given the large number (4900) of unique panels generated for this study and the concordance with previous results generated from real data [33], it is reasonable to believe that the findings are generalizable. Moreover, as already explained, the assumptions made in the generation of the data are grounded in educational measurement and standard settings theories and findings in practice [12, 29, 32, 35, 47, 67]. Note that as expected from a simulation study, this study measures the quality of a model rather than analysing any observed data [22].

Further research is needed to identify the impact of other features of modified Angoff methods on cut-score precision, as well as repeating this study using modified assumptions.

Conclusions and practical advice

This study demonstrates that Angoff is not only a reliable method, as previously suggested, but it is also a valid method for setting assessment standards. There are three main practice points that emerge from this study: (1) a panel of about fifteen judges provides a reliable and valid cut-score for a test consisting of 80 items or above. However, when the number of items is fewer than 15, it is recommended that no fewer than 20 judges are used, with the number of judges increasing as the number of items declines; (2) panels should include judges with mixed levels of expertise unless there is clear evidence that the experts among the panellists are not more stringent than the non-experts; and (3) without providing additional test data to the judges, a second round of Angoff is redundant.

Lack of data of the true ability of examinees remains a major hurdle in the field of standard setting. Thus research utilising simulation methods may enhance our understanding about the validity and applicability of a range of standard setting methods beyond what was demonstrated in this study.

Ethics approval and consent to participate

This study used computer generated data i.e. simulated data, hence consent to participate was not necessary. Thus, ethics approval was not required.

Availability of supporting data

Data used in this study may be available by request. Please contact A/Prof Boaz Shulruf b.shulruf@unsw.edu.au.

Competing interests

This research has no financial implication for any of the authors, no funding was given for undertaking this research. This research has no impact on any financial and/or non-financial interest of any of the authors or their affiliated organisations.

Authors' contributions

BS initiated and designed the study, developed the conceptual framework, undertook the statistical analysis and has contributed to the interpretation of the results and the writing of the manuscript. PP contributed to the development of concept and study design, contributed to the analysis and interpretation of the data and contributed to the writing of the manuscript. TW contributed to the development of concept and study design, contributed to the analysis and interpretation of the data and contributed to the writing of the manuscript. JW contributed to the development of concept and study design, contributed to the analysis and interpretation of the data and contributed to the writing of the manuscript. PJ contributed to the development of concept and study design, contributed to the analysis and interpretation of the data and contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Author details

[1]University of New South Wales, Sydney, Australia. [2]University of Auckland, Auckland, New Zealand. [3]Otago University Christchurch, Christchurch, New Zealand.

Received: 12 August 2015 Accepted: 28 April 2016
Published online: 04 May 2016

References

1. Angoff W. Test reliability and effective test length. Psychometrika. 1953;18(1):1–14. doi:10.1007/BF02289023.
2. Angoff W. Scales, norms, and equivalent scores. In: Thorndike R, editor. Educational measurement. 2nd ed. Washington, DC: American council on education; 1971. p. 508–600.
3. Arce A, Wang Z. Applying Rasch model and generalizability theory to study Modified-Angoff cut scores. Int J Test. 2012;12(1):44–60. doi:10.1080/15305058.2011.614366.
4. Behuniak P, Archambault F, Gable R. Angoff and Nedelsky standard setting procedures: implications for the validity of proficiency test score interpretation. Educ Psychol Meas. 1982;42(1):247–55. doi:10.1177/0013164482421031.
5. Ben-David M. AMEE Guide No. 18: Standard setting in student assessment. Med Teach. 2000;22(2):120–30. doi:10.1080/01421590078526.
6. Boursicot K, Roberts T. How to set up an OSCE. Clin Teach. 2005;2(1):16–20. doi:10.1111/j.1743-498X.2005.00053.x.
7. Brandon P. Conclusions about frequently studied modified angoff standard-setting topics. Appl Meas Educ. 2004;17(1):59–88. doi:10.1207/s15324818ame1701_4.
8. Buckendahl CW, Smith RW, Impara JC, Plake BS. A comparison of Angoff and bookmark standard setting methods. J Educ Meas. 2002;39(3):253–63. doi:10.1111/j.1745-3984.2002.tb01177.x.
9. Burns K. Classical reliability: using generalizability theory to assess dependability. Res Nurs Health. 1998;21(1):83–90.
10. Busch J, Jaeger R. Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. J Educ Meas. 1990;27(2):145–63. doi:10.1111/j.1745-3984.1990.tb00739.x.
11. Chang L. Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. Appl Meas Educ. 1999;12(2):151–65.
12. Cizek G. Setting performance standards: foundations, methods, and innovations. 2nd ed. London: Routledge; 2012.
13. Cizek G, Bunch M. The Angoff Method and Angoff Variations. In: Cizek G, Bunch M, editors. Standard Setting. Thousand Oaks, California: SAGE Publications, Inc.; (2007a). p. 81–96.
14. Cizek G, Bunch M. Standard setting: a guide to establishing and evaluating performance standards on tests. London: Sage Pubns; 2007b.
15. Clauser B, Harik P, Margolis M, McManus I, Mollon J, Chis L, Williams, S. An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. Appl Meas Educ. 2009;22(1):1–21. doi:10.1080/08957340802558318.
16. Clauser B, Mee J, Baldwin S, Margolis M, Dillon G. Judges' use of examinee performance data in an angoff standard-setting exercise for a medical licensing examination: an experimental study. J Educ Meas. 2009;46(4):390–407. doi:10.1111/j.1745-3984.2009.00089.x.
17. Clauser J, Margolis M, Clauser B. An examination of the replicability of angoff standard setting results within a generalizability theory framework. J Educ Meas. 2014;51(2):127–40. doi:10.1111/jedm.12038.
18. Cleemput I, Kind P, Kesteloot K. Re-scaling social preference data: implications for modelling. Eur J Health Econ. 2004;5(4):290–8. doi:10.1007/s10198-004-0242-5.
19. Cusimano M. Standard setting in medical education. Acad Med. 1996;71(10):S112–120.
20. Cusimano M, Rothman A. The effect of incorporating normative data into a criterion-referenced standard setting in medical education. Acad Med. 2003;78(10):S88–90.
21. Davis-Becker S, Buckendahl C, Gerrow J. Evaluating the bookmark standard setting method: the impact of random item ordering. Int J Test. 2011;11(1):24–37. doi:10.1080/15305058.2010.501536.
22. Dorans N. Simulate to understand models, Not nature. ETS Res Rep Ser. 2014;2014(2):1–9. doi:10.1002/ets2.12013.
23. Engelhard G. Examining rater errors in the assessment of written composition with a many-faceted Rasch model. J Educ Meas. 1994;31(2):93–112.
24. Fan X. Designing Simulation Studies. In: Cooper H, editors. APA Handbook of Research Methods in Psychology: Quantitative, Qualitative, Neuropsychological, and Biological (Vol. 2). Washington, D.C.: American Psychological Association; 2012.
25. Ferdous A, Plake B. Item selection strategy for reducing the number of items rated in an Angoff standard setting study. Educ Psychol Meas. 2007;67(2):193–206. doi:10.1177/0013164406288160.
26. Fowell SL, Fewtrell R, McLaughlin PJ. Estimating the minimum number of judges required for test-centred standard setting on written assessments. Do discussion and iteration have an influence? Adv Health Sci Educ. 2008;13(1):11–24. doi:10.1007/s10459-006-9027-1.
27. Garson D. Creating simulated datasets. Asheboro: North Carolina state University and G. David Garson and Statistical Associates Publishing; 2012.
28. George S, Haque S, Oyebode F. Standard setting: comparison of two methods. BMC Med Educ. 2006;6(1):46.
29. Gipps C. Assessment paradigms. In: Gipps C, editor. Beyond testing: towards a theory of educational assessment. London: The Falmer Press; 1994. p. 1–18.
30. Harvill L. Standard error of measurement. Educ Meas. 1991;10(2):33–41. doi:10.1111/j.1745-3992.1991.tb00195.x.
31. Houston W, Raymond M, Svec J. Adjustments for rater effects in performance assessment. Appl Psychol Meas. 1991;15(4):409–21. doi:10.1177/014662169101500411.
32. Hurtz G, Auerbach MA. A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. Educ Psychol Meas. 2003;63(4):584–601. doi:10.1177/0013164403251284.
33. Hurtz G, Hertz N. How many raters should be used for establishing cutoff scores with the Angoff method? A generalizability theory study. Educ Psychol Meas. 1999;59(6):885–97. doi:10.1177/00131649921970233.
34. Hurtz G, Patrick J. Innovations in measuring rater accuracy in standard setting: assessing "Fit" to item characteristic curves. Appl Meas Educ. 2009;22(2):120–43. doi:10.1080/08957340902754601.
35. Hutchison D. On the conceptualisation of measurement error. Oxford Rev Educ. 2008;34(4):443–60. doi:10.1080/03054980701695662.
36. Jaeger R. Selection of judges for standard-setting. Educ Meas. 1991;10(2):3–14. doi:10.1111/j.1745-3992.1991.tb00185.x.
37. Jalili M, Hejri S, Norcini J. Comparison of two methods of standard setting: the performance of the three-level Angoff method. Med Educ. 2011;45(12):1199–208. doi:10.1111/j.1365-2923.2011.04073.x.
38. Kaliski PK, Wind SA, Engelhard G, Morgan DL, Plake BS, Reshetar RA. Using the many-faceted Rasch model to evaluate standard setting judgments: an illustration with the advanced placement environmental science exam. Educ Psychol Meas. 2013;73(3):386–411. doi:10.1177/0013164412468448.
39. Kramer A, Muijtjens A, Jansen K, Düsman H, Tan L, Van Der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. Med Educ. 2003;37(2):132–9. doi:10.1046/j.1365-2923.2003.01429.x.
40. Liao SC, Hunt EA, Chen W. Comparison between inter-rater reliability and inter-rater agreement in performance assessment. Ann Acad Med Singapore. 2010;39:613–8.
41. Livingston SA, Zieky MJ. Passing scores: manual for setting standards of performance eonducational and occupational tests. Princeton Princeton: Educational Testing Service; 1982.
42. MacCann RG, Stanley G. The use of Rasch modeling to improve standard setting. Pract Assess Res Eval. 2006;11(2):1–17. http://pareonline.net/genpare.asp?wh=0&abt=11.
43. Margolis M, Clauser B. The impact of examinee performance information on judges' cut scores in modified Angoff standard-setting exercises. Educ Meas. 2014;33(1):15–22. doi:10.1111/emip.12025.
44. Maurer T, Alexander R, Callahan C, Bailey J, Dambrot F. Methodological and psychometric issues in setting cutoff scores using the Angoff method. Person Psychol. 1991;44(2):235–62. doi:10.1111/j.1744-6570.1991.tb00958.x.
45. McKinley D, Norcini J. How to set standards on performance-based examinations: AMEE Guide No. 85. Med Teach. 2014;36(2):97–110. doi:10.3109/0142159X.2013.853119.
46. Mehrens W, Popham J. How to evaluate the legal defensibility of high-stakes tests. Appl Meas Educ. 1992;5(3):265.
47. Nichols P, Twing J, Mueller CD, O'Malley K. Standard-setting methods as measurement processes. Educ Meas. 2010;29(1):14–24. doi:10.1111/j.1745-3992.2009.00166.x.
48. Norcini J, Lipner R, Langdon L, Strecker C. A comparison of three variations on a standard-setting method. J Educ Meas. 1987;24(1):56–64. doi:10.1111/j.1745-3984.1987.tb00261.x.
49. Page G, Bordage G. The Medical Council of Canada's key features project: A more valid written examination of clinical decision-making skills. Acad Med. 1995;70(2):104–10.
50. Pant H, Rupp A, Tiffin-Richards S, Köller O. Validity issues in standard-setting studies. Stud Educ Eval. 2009;35(2–3):95–101. doi: http://dx.doi.org/10.1016/j.stueduc.2009.10.008.

51. Peterson C, Schulz EM, Engelhard Jr G. Reliability and validity of bookmark-based methods for standard setting: comparisons to Angoff-based methods in the National Assessment of Educational Progress. Educ Meas. 2011;30(2):3–14. doi:10.1111/j.1745-3992.2011.00200.x.
52. Raymond M, Luciw-Dubas U. The second time around: accounting for retest effects on oral examinations. Eval Health Prof. 2010;33(3):386–403. doi:10.1177/0163278710374855.
53. Reckase M. A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. Educ Meas. 2006;25(2):4–18. doi:10.1111/j.1745-3992.2006.00052.x.
54. Ricker K. Setting cut-scores: a critical review of the Angoff and modified Angoff methods. Alberta J Educ Res. 2006;52(1):53–64.
55. Schoonheim-Klein M, Muijtjens A, Habets L, Manogue M, van der Vleuten C, van der Velden U. Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard setting methods. Eur J Dent Educ. 2009;13(3):162–71. doi:10.1111/j.1600-0579.2008.00568.x.
56. Schuwirth L, Colliver J, Gruppen L, Kreiter C, Mennin S, Onishi H, Wagner-Menghin, M. Research in assessment: consensus statement and recommendations from the Ottawa 2010 conference. Med Teach. 2011;33(3):224–33. doi:10.3109/0142159X.2011.551558.
57. Schuwirth L, van der Vleuten C. A plea for new psychometric models in educational assessment. Med Educ. 2006;40(4):296–300. doi:10.1111/j.1365-2929.2006.02405.x.
58. Shulruf B, Poole P, Jones P, Wilkinson T. The Objective Borderline Method (OBM): a probabilistic method for standard setting. Assess Eval High Educ. 2014. doi:10.1080/02602938.2014.918088.
59. Skorupski W. Understanding the cognitive processes of standard setting panelists. In: Zieky MJ, editor. Setting performance standards: foundations, mathods, and innovations. 2nd ed. London: Routledge; 2012. p. 135–47.
60. Swanlund A, Smith E. Developing examinations that use equal raw scores for cut scores. J Appl Meas. 2010;11(4):432–42.
61. Tannenbaum R, Kannan P. Consistency of angoff-based standard-setting judgments: are item judgments and passing scores replicable across different panels of experts? Educ Assess. 2015;20(1):66–78. doi:10.1080/10627197.2015.997619.
62. Tavakol M, Dennick R. Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE Guide No. 72. Med Teach. 2013;35(1):e838–48. doi:10.3109/0142159X.2012.737488.
63. Van Nijlen D, Janssen R. Modeling judgments in the Angoff and contrasting-groups method of standard setting. J Educ Meas. 2008;45(1):45–63. doi:10.1111/j.1745-3984.2007.00051.x.
64. Verheggen M, Muijtjens A, Van Os J, Schuwirth L. Is an Angoff standard an indication of minimal competence of examinees or of judges? Adv Health Sci Educ. 2008;13(2):203–11. doi:10.1007/s10459-006-9035-1.
65. Verhoeven B, Van der Steeg A, Scherpbier A, Muijtjens A, Verwijnen G, Van Der Vleuten C. Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. Med Educ. 1999;33(11):832–7. doi:10.1046/j.1365-2923.1999.00487.x.
66. Verhoeven B, Verwijnen G, Muijtjens A, Scherpbier A, van der Vleuten C. Panel expertise for an Angoff standard setting procedure in progress testing: item writers compared to recently graduated students. Med Educ. 2002;36(9):860–7. doi:10.1046/j.1365-2923.2002.01301.x.
67. Viswanathan M. What causes measurement error? New York: Sage; 2005.
68. Wang N. Use of the Rasch IRT model in standard setting: an item-mapping method. J Educ Meas. 2003;40(3):231–53. doi:10.1111/j.1745-3984.2003.tb01106.x.
69. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357(9260):945–9. doi:10.1016/S0140-6736(00)04221-5.
70. Wayne D, Cohen E, Makoul G, McGaghie W. The impact of judge selection on standard setting for a patient survey of physician communication skills. Acad Med. 2008;83(10):S17–20. doi:10.1097/ACM.1090b1013e318183e318187bd.
71. Wood T, Humphrey-Murto S, Norman G. Standard setting in a small scale OSCE: a comparison of the modified borderline-group method and the borderline regression method. Adv Health Sci Educ. 2006;11(2):115–22. doi:10.1007/s10459-005-7853-1.
72. Wyse A, Reckase M. Examining rounding rules in Angoff-type standard-setting methods. Educ Psychol Meas. 2012;72(2):224–44. doi:10.1177/0013164411413572.
73. Yelle LE. The learning curve: historical review and comprehensive survey. Decis Sci. 1979;10(2):302–28. doi:10.1111/j.1540-5915.1979.tb00026.x.
74. Yudkowsky R, Downing S, Popescu M. Setting standards for performance tests: a pilot study of a three-level Angoff method. Acad Med. 2008;83(10):S13–6. doi:10.1097/ACM.1090b1013e318183c318683.
75. Yudkowsky R, Downing S, Wirth S. Simpler standards for local performance examinations: the yes/no Angoff and Whole-Test Ebel. Teach Learn Med. 2008;20(3):212–7. doi:10.1080/10401330802199450.
76. Zieky MJ. So much has changed. In: Cizek G, editor. Setting performance standards: foundations, mathods, and innovations. 2nd ed. London: Routledge; 2012. p. 15–32.