Informed Health Choices

Claim evaluation tools

Measuring ability to assess claims about treatment effects: establishment of a standard for passing and mastery

Alun Davies et al.
Working paper, 9 January 2017

Colophon

Title Measuring ability to assess claims about treatment effects: establishment of a standard for passing and mastery

Authors
Davies, Alun
Gerrity, Martha
Nordheim, Lena
Okebukola Peter
Opiyo, Newton
Sharples, Jonathan
Wilson, Helen
Wysonge, Charles
Austvoll-Dahlgren, Astrid
Oxman, Andrew David

Corresponding author(s)
Oxman, Andrew
oxman@online.no
Norwegian Institute of Public Health
PO Box 4404, Nydalen
N-0403 Oslo, Norway

Keywords critical thinking, critical appraisal, informed health choices, evidence-based healthcare, health literacy, outcome measurement, multiple choice questions, absolute criterion-referenced standard

Citation Davies A, Gerrity M, Nordheim LV, Opiyo N, Okebukola PO, Sharples J, Wilson H, Wiysonge C, Austvoll-Dahlgren A, Oxman AD. Measuring ability to assess claims about treatment effects: establishment of a standard for passing and mastery. IHC Working Paper, 2017. ISBN 978-82-8082-802-6.
http://www.informedhealthchoices.org/wp-content/uploads/2016/08/Claim-cut-off-IHC-Working-Paper-2017-01-09.pdf

Article category
☐ About Informed Health Choices
☐ Key concepts and glossary
☐ Learning resources
☐ Systematic reviews
☐ Development and evaluation of learning resources
☑ **Claim evaluation tools**
☐ Editorials and commentaries
☐ Grant applications

Abstract

Background: We selected two sub-sets of multiple-choice questions from the Claim Evaluation Tools database to create the Informed Health Choices (IHC) primary school test and the IHC podcast test to evaluate the effectiveness of learning resources for year-5 (10- to 12-year-old) children in Uganda and a podcast for their parents. The learning resources and the podcast focus on Key Concepts that people need to understand and be able to apply to assess claims about the effects of a treatment (any action intended to improve health) and to make informed health choices. The primary school test was used to measure the ability of the children to apply 12 Key Concepts that are taught in the learning resources. The podcast test was used to measure the ability of the parents to apply nine Key Concepts that are covered by the podcast. Both tests included two multiple-choice questions (MCQs) for each concept. The tests had eight concepts (16 MCQs) in common.

Objective: The objectives of this study were to determine cut-off scores for passing (having at least a borderline ability to apply the concepts) and mastery (having mastered the concepts) for the IHC primary school and podcast tests.

Methods: Eight people judged the likelihood that someone with a borderline ability and someone who had mastered the concepts would answer each MCQ correctly. They determined cut-off scores by summing up the probability of answering each MCQ correctly. They were provided with instructions based on a combination of two methods for setting standards for performance on educational tests (Nedelsky's and Angoff's). In two groups of four judges, the instructions were discussed, there was a practice round, and another discussion. After that, the judges made their assessments independently. These were summarised for each group of judges and the judges reached a consensus.

Results: The judges agreed that for the primary school test, 13 or more questions out of 24 needed to be answered correctly to pass and 20 or more questions to demonstrate mastery. For the podcast test, 11 or more questions out of 18 needed to be answered correctly to pass and 15 or more questions for mastery.

Conclusions: We found that it was possible to quickly reach a consensus during an online meeting and to agree on cut-off scores using a combination of Nedelsky's and Angoff's methods.

Background

The Informed Health Choices (IHC) project has identified Key Concepts that people need to understand and apply to assess claims about the effects of a treatment (any action intended to improve health) and to make informed health choices.[1] This includes concepts about claims and whether they are justified, about comparisons and whether they are fair and reliable, and about using evidence to make informed choices. The Claim Evaluation Tools database contains multiple-choice questions (MCQs) that can be used to measure someone's ability to apply those concepts.[2-4] The MCQs can be used by an individual for self-assessment, by teachers to assess learners' abilities, and by researchers to evaluate learning resources or to map people's abilities.

We used MCQs from the Claim Evaluation Tools database to create an outcome measure to be used in an evaluation of the effects of the IHC primary school resources. These learning resources were designed to teach 10- to 12-year-old children 12 of the Key Concepts (Table 1). We evaluated them in a randomised trial in Uganda.[5] In a linked trial, we evaluated the IHC podcast, which was designed to inform parents of those children about nine of the Key Concepts.

Table 1. *Key Concepts addressed by the primary school and podcast tests*

Primary school test	Key Concepts	Podcast test
	Claims: Are they justified?	
1	Treatments may be harmful	1
2	Personal experiences or anecdotes (stories) are an unreliable basis for assessing the effects of most treatments	2
	A treatment outcome may be associated with a treatment, but not caused by the treatment	3
3	Widely used treatments or treatments that have been used for a long time are not necessarily beneficial or safe	4
4	New, brand-named, or more expensive treatments may not be better than available alternatives	
5	Opinions of experts or authorities do not alone provide a reliable basis for deciding on the benefits and harms of treatments	5
6	Conflicting interests may result in misleading claims about the effects of treatments	
	Comparisons: Are they fair and reliable?	
7	Evaluating the effects of treatments requires appropriate comparisons	6
8	Apart from the treatments being compared, the comparison groups need to be similar (i.e. 'like needs to be compared with like')	7
9	If possible, people should not know which of the treatments being compared they are receiving	
10	Small studies in which few outcome events occur are usually not informative and the results may be misleading	
11	The results of single comparisons of treatments can be misleading	8
	Choices: Making informed health choices	
12	Treatments usually have beneficial and harmful effects	9

As can be seen in Table 1, the primary school resources and the podcast had eight Key Concepts in common. There were two MCQs for each Key Concept in the two outcome measures. In this paper, we will refer to the 24 MCQs used in the primary school trial as the primary school test, and the 18 MCQs in the podcast trial as the podcast test.

It is difficult to interpret average differences in scores for a test or other continuous (or count) outcome measures.[7] Doing so requires a basis for judging the importance of any differences. In addition, it requires examining the distribution of the scores. For example, a small average difference in test scores might be due to most students doing a little bit better or to a few students doing a lot better in a comparison of two groups of learners.

The difference in the proportion of people who have a passing score is more meaningful and easier to interpret than an average difference in test scores. In this context, passing means:

- having a basic understanding of the concepts and how to apply them
- not needing to repeat the lessons, listen to the podcast again, or receive some other additional or alternative instruction
- being ready to go on to other lessons or another podcast that reinforce learning of the same concepts and introduce new concepts

Determining the proportion of people who pass requires determining a cut-off score, above which someone passes and below which someone does not, or in this context:

- those above the cut-off have a basic understanding of the concepts and are able to apply them, whereas the those below the cut-off do not
- those below the cut-off need to repeat the lessons, listen again to the podcast or receive some other additional or alternative instruction, whereas those above the cut-off do not
- those above the cut-off are ready to go on to other lessons or another podcast, which will reinforce learning of the same concepts and introduce new concepts, whereas those below the cut-off are not

Options for determining a cut-off and the reason for using an absolute (or criterion referenced) standard are summarised in Table 2.

Table 2. *Options for determining a cut-off*

Option	Comments
Core concepts	This approach bases passing on answering correctly all the questions that address concepts that are considered core or necessary. Individuals making one mistake would fail. This approach ignores concepts that are not considered "core" and it ignores how difficult the questions are.
Relative standards (norm referenced)	This approach is based on a comparison among the performances of the individuals taking the test. A set proportion of candidates fails. Using relative standards does not make sense in the context of evaluating the effectiveness of an educational intervention that has specific measurable objectives.
Absolute standards (criterion referenced)	This approach is based on how much an individual knows and can apply. Individuals pass or fail depending on whether they meet a specified criterion.
Mixed or compromise methods	This approach combines the use of a relative and an absolute approach. It requires a norm (setting a proportion that fail), which does not make sense in the context of evaluating the effectiveness of an educational intervention.

The Nedelsky, Angoff, and Ebel methods (and modifications of these) are used for determining an absolute standard.[8] All three rely on the concept of individuals who are on the borderline of passing or failing and expert judges. With Nedelsky's method,[9] the judges eliminate response options that a borderline individual would be able to eliminate. The chances of getting each question correct is then equal to one divided by the number of remaining response options; e.g. if there are two remaining response options (one of which is the correct option), the chances of a borderline individual answering the question correct is 1/2 or 50%. The cut-off score is then determined by adding up the probabilities for all the questions.

Angoff's method, which is one of the most widely used methods, is similar to Nedelsky's,[10] but the judges assess the difficulty of each question as a whole, instead of making judgements about each response option.

Ebel's method is similar to Angoff's,[11] but judges are asked to make two judgements: one about the relevance of each question and one about the difficulty of each question. They then judge the difficulty of the questions in each cell of a matrix (of levels of relevance by levels of difficulty).

We used a combination of Nedelsky's and Angoff's methods. The judges started with Nedelsky's method, then increased or decreased the assigned probability for each question based on an overall assessment. This gave the judges a logical

approach to making an initial judgement about the difficulty of each question. It then allowed them to adjust for uncertainty about the number of response options a borderline individual would eliminate, the difficulty of the stem (scenario) for the question, the difficulty of the concept, and anything else that might make a question more or less difficult.

For each of the above methods there are five steps:

1. select the judges
2. define "borderline" knowledge and ability
3. train the judges in the use of the method
4. collect their judgements
5. combine the judgement to choose a passing score

Methods

Selection of the judges

The judges must be qualified to decide what level of the knowledge or skills measured by the test is necessary to conclude that an individual:

- has at least a basic understanding of the concepts and ability to apply them
- does not need to repeat the lessons, listen to the podcast again, or receive some other additional or alternative instruction
- is ready to go on to other lessons or another podcast, which will reinforce learning of the same concepts and introduce new concepts

Because of the nature of these tests and what we are trying to measure, one of us (ADO) purposively selected and recruited two types of judges in May 2016: health researchers and people who teach evidence-informed decision making, and education researchers with experience evaluating interventions to teach critical thinking skills (Table 3). In addition, teachers who participated in pilot testing of the IHC primary school resources reviewed the judgements that were made to ensure they were appropriate for the target audience and the context.

Table 3. *Judges*

	Sex	Country	Background
Judge 1	M	UK / Kenya	Teacher / health and education researcher
Judge 2	M	Cameroon / South Africa	Health researcher
Judge 3	M	UK	Education researcher
Judge 4	F	USA	Health researcher with PhD in education
Judge 5	F	UK	Teacher / education researcher
Judge 6	F	Norway	Health and education researcher
Judge 7	M	Kenya / Norway	Health researcher
Judge 8	M	Nigeria / USA	Health researcher

Definition of borderline knowledge and ability

An individual with borderline knowledge and ability is someone who may or may not have a basic understanding of the concepts and ability to apply them, may or may not need additional or alternative instruction, and may or may not be ready to go on to other lessons or another podcast. We created personas who are characteristic of people with borderline knowledge and ability and of people who clearly has mastered the concepts (Appendix 1). These were used to communicate and help the judges to envisage these people.

Training of the judges

We provided the judges with instructions (Appendix 1) and discussed these with them before they started making their judgements.

The judges took the combined test (with 26 MCQs) before making judgements about the difficulty of the questions. We then gave them the right answers so that they had these when they made their judgements. In this way, they got a better sense of how difficult the questions were than if they were given the answers before taking the test themselves.

The judges participated in a practice round with MCQs that had different degrees of difficulty before they made their individual judgements. This exercise allowed them to discuss what makes a question difficult or easy. It also made them aware of their tendencies to be more or less pessimistic about the probability of a borderline test-taker answering questions correctly in comparison with the other judges, and to ameliorate their judgments accordingly when they assessed the full set of MCQs.

Collecting the judgements

The judges independently judged all the MCQs. Since they were not judging relevance (just difficulty) they could judge all the MCQs for the 13 concepts covered in either test without having to repeat their judgements for the 16 MCQs that are used in both.

Combining the judgements to choose a passing score

One of us (ADO) calculated the mean and median for each MCQ and for the cut-off score. He presented both these statistics and the range to the judges. He also showed the judges the difficulty of the MCQs based on the Rasch analysis - after they had made their judgements.[3] He moderated an online discussion where the judges discussed any discrepancies in the relative difficulty of the MCQs between their judgements and the results of the Rasch analysis. He then asked them to discuss each MCQ and reach a consensus. Because it was not possible to schedule a time that was convenient for all the judges, there were two separate

discussion with four judges in each group. The two groups were shown the judgements for both groups, and the second group was shown the consensus judgement for the first group.

We used a modified nominal group approach to reach a consensus.[12] We first showed everybody all the judgements for each MCQ. We then invited people from each end of the range to provide the reasons for their judgements, and then invited others to comment. After the final cut-off score was decided, we checked to make sure that all the judges agreed with the cut-off scores, and adjusted them as necessary, based on a consensus of all the judges.

Determining a cut-off score for mastery

We also asked the judges to make the same set of judgements to set a second cut-off for a score that indicates mastery of the concepts, using the same approach. This cut-off is the minimum score that they would expect for an individual who clearly has a basic understanding of the concepts and ability to apply them, does not need additional or alternative instruction, and is ready to go on to other lessons or another podcast, which will reinforce learning of the same concepts and introduce new concepts.

Results

After the pilot, the judges agreed on the following guidance, which they found helpful:

- *There is always a chance of reading problems – as a general rule always go down at least 10% for reading errors for both borderline test takers and mastery and 20% for borderline test takers when the initial probability of a correct answer is 100% for both borderline test takers and mastery.*
- *The difficulty of some of these questions may vary from setting to setting. We did not find this in the Rasch analysis, but we only tested this to a limited extent. We are testing the questions in other countries. For the purposes of this exercise, it is best to keep in mind the context in which we are using these tests and these cut-off scores.*

Nonetheless, the judges found it difficult to assess the difficulty of each MCQ. When we discussed the reasoning that they used, we found that different judges had different reasons for their judgements, and each judge tended to apply the same reasoning across MCQs. Because most of the judges were not biased towards over or under-estimating the difficulty of the MCQs, there was less variation when their judgements across the questions were summarised.

Thus, there was substantial disagreement in the judges' independent judgements about how difficult each MCQ was (Appendix 2). However, there was less disagreement when the probabilities for each MCQ were added up to determine the cut-offs and the judges quickly reached a consensus about both the difficulty of each MCQ and the cut offs (Table 4). Each group of judges met online July 11, 2016. The group of judges that met later in the day agreed with the consensus that was reached by the first group.

For the primary school test, 13 or more questions out of 24 need to be answered correctly to pass and 20 or more questions out of 24 need to be answered correctly to demonstrate mastery.

For the podcast test, 11 or more questions out of 18 need to be answered correctly to pass and 15 or more questions out of 18 need to be answered correctly to demonstrate mastery.

The teachers who participated in pilot testing of the IHC primary school resources felt that these cut-offs were appropriate for the target audience and the context.

Table 4. *Individual and consensus summary judgements*

		Group 1							Group 2							Both groups				
Judges:		**1**	**2**	**3**	**4**	**Avg**	**Med**	**Consensus**	**5**	**6**	**7**	**8**	**Avg**	**Med**	**Consensus**	**Avg**	**Med**	**Min**	**Max**	**Consensus**
Primary school test (Children)																				
Rasch																			9.89	
Chance																			9.25	
Out of 24	Pass score	12.4	14.6	12.7	12.2	13.0	12.8	**12.7**	10.9	15.1	13.5	7.7	11.8	11.7	**11.7**	12.4	12.4	7.7	15.1	**13 out of 24**
Out of 24	Master score	18.1	20.7	21.0	17.8	19.6	20.5	**20.1**	19.7	21.0	21.5	10.1	18.2	20.4	**20.4**	18.9	20.6	10.1	21.5	**20 out of 24**
Podcast test (Adults)																				
Rasch																			7.0	
Chance																			6.7	
Out of 18	Pass score	8.6	10.5	9.1	9.0	9.3	9.3	**9.1**	8.1	10.7	9.5	5.7	8.5	8.5	**8.5**	8.9	9.0	5.7	10.7	**11 out of 18**
Out of 18	Master score	12.7	15.3	15.7	13.3	14.4	15.1	**14.7**	14.2	15.5	15.7	7.5	13.3	14.9	**14.9**	13.9	15.2	7.5	15.7	**15 out of 18**

Judges = 1 to 8; Avg = average; Med = median; Min = minimum; Max = maximum

Rasch = expected score based on difficulty of each question from Rasch analysis (proportion of participants who answered each question correctly)[3]

Chance = expected score by chance alone (guessing)

Discussion

There was substantial variation in the judges' independent assessments of the difficulty of each MCQ. The variability may have been partly due to differences in the backgrounds of the judges, who came from different disciplines (health and education) and countries (Table 3). This is consistent with the findings of multiple studies showing that judges struggle with making the required judgements.[13] However, the judges quickly reached agreement. It helped to have the judges record their reasoning when they made their judgements so that they could refer to this when they discussed their judgements and reached a consensus.

We provided the judges with performance data - the proportion of test takers who answered each question correctly (based on data from a Rasch analysis)[3] - after they made their independent judgements. Studies about the effects of providing judges with empirical data have shown that this can change the cut score.[13-17] Some authors consider it important to provide judges with performance data, whereas others are concerned about the potential for judges to rely too heavily on the performance data, and some consider the need to provide judges with empirical data to be a serious flaw in the Angoff procedure.[13] We found that providing judges with performance data after they made their independent judges - together with a summary of their judgements - was helpful and this might help to address concerns about overreliance on performance data while, at the same time, enabling judges to use these data to inform their final judgements and reach a consensus. Studies of the impact of judges iteratively judging items first without and then with performance data have shown that this increases the internal consistency of the judgments, bringing them into closer correspondence with the empirical probabilities, and often leads to important changes in the estimated cut score.[13-17]

Changes in the instructions given to the judges can affect the extent to which they rely on performance data.[16] In our study, the judges were told that while the performance data provided an indication of the relative difficulty of the questions, it did not provide an accurate indication of the probability of a borderline test-taker answering a question correctly, since most of the data in the Rasch analysis came from people who were unlikely to pass.

Studies comparing the Angoff and Nedelsky methods have found that they can result in different cut-off scores, and that the restricted nature of the Nedelsky method (focusing on the response options only) may limit its usefulness.[18-20] Nonetheless, strengths of the Nedelsky method are lower intrajudge incon-

sistency, which is attributable to focusing on response options and making multiple decisions.[20] It has been suggested that combining the Nedelsky and Angoff methods would make a stronger standard-setting procedure.[20] However, we are not aware of empirical evaluations of this approach. We found that using Nedelsky's method provided a helpful starting point for making judgements about the difficulty of questions and that recording these judgements (Appendix 2) helped to resolve disagreements and reach a consensus about the overall difficulty of each question.

Empirical studies have compared instructing judges to consider whether a single minimally competent (borderline) candidate would or would not answer each question correctly (using a yes/no procedure) to instructing them to judge the proportion of minimally competent test takers who would answer each question correctly. Although the results are equivocal, they suggest that these two approaches result in similar cut-offs.[21] We did not use a yes/no procedure. However, drawing on experience from design methods, we created personas to help the judges to visualise typical test takers with borderline knowledge and ability and ones who clearly had mastered the concepts. We found this helpful.

Empirical studies have found that judgements made using the Angoff method are reproducible,[22] but also that there can be variability in cut-off scores set by different groups of judges.[23] It is uncertain to what extent small group discussion among the judges leads to more reliable and valid cut-off scores.[24] Although we did not plan on having two groups of judges with four judges in each group, we found that this worked well. It allowed more time for each judge to participate in the discussions when a consensus was reached, facilitated reaching a consensus in each group in less than one hour, and allowed us to compare the consensus of the two groups of judges. It should, however, be noted that this is lower than the number of judges that is considered by others to be an acceptable minimum (10).[22,25]

The Angoff method, which we used, is one of the best known and most widely used methods of standard setting, although other methods are advocated and used.[26,27] The validity of all of these methods has been questioned.[19] Overall, the Angoff method and modifications of the Angoff method such as we have used meet nine of the ten criteria proposed by Berk for assessing methods for setting cut-off scores.[21] The main strength of these methods is theirts simplicity, whereas the main drawback is the cognitive burden on the judges, particularly if there are many questions that must be assessed and if multiple iterations are needed. In this study, although the judges struggled with the judgements they were asked to make, there were not many questions, the judges reached agreement quickly after a single round of independent judgements, and we were able to quickly establish cut-off scores for passing and for mastery.

Conclusion

Setting a cut-off for passing or mastery is challenging. However, we found that it was possible to quickly reach a consensus during a one hour online meeting (after the judges independently assessed each question) and to agree on cut-off scores that all eight judges agreed were sensible. We found it helpful to use a combination of Nedelsky's and Angoff's methods, to agree on some general guidance following the pilot, for the individual judges to record the reasons for their judgements, and to use a consensus among small groups of judges.

References

1. Austvoll-Dahlgren A, Oxman AD, Chalmers I, Nsangi A, Glenton C, Lewin S, et al. Key concepts that people need to understand to assess claims about treatment effects. J Evid Based Med 2015; 8:112-25.
2. Austvoll-Dahlgren A, Semakula D, Nsangi A, et al. The development of the "Claim Evaluation Tools": assessing critical thinking about effects. BMJ Open, in press.
3. Austvoll-Dahlgren A, Guttersrud Ø, Semakula D, Nsangi A, Oxman AD. Measuring ability to assess claims about treatment effects: A latent trait analysis of the Claim Evaluation Tools using Rasch modelling. BMJ Open, in press.
4. Nsangi A, Semakula D, Austvoll-Dahlgren A, Guttersrud Ø, Oxman AD. Measuring ability to assess claims about treatment effects in Luganda and English: a latent trait analysis of two Claim Evaluation short versions. In preparation.
5. Nsangi A, Semakula D, Oxman AD, et al. Does the use of the Informed Healthcare Choices (IHC) primary school resources improve the ability of year-5 children in Uganda to assess the trustworthiness of claims about the effects of treatments: protocol for a cluster-randomised trial. Trials, in press.
6. Semakula DN, Nsangi A, Oxman AD, et al. Does the use of an educational podcast improve the ability of parents of primary school children to assess the trustworthiness of claims about the effects of treatments: Protocol for a randomised trial? Trials, in press.
7. Guyatt GH, Thorlund K, Oxman AD, Walter SD, Patrick D, Furukawa TA, et al. GRADE guidelines: 13. Preparing summary of findings tables – continuous outcomes. J Clin Epidemiol 2013; 66:173-83.
8. Livingston SA, Zieky MJ. Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Princeton, NJ: Educational Testing Service, 1982.
9. Nedelsky L. Absolute grading standards for objective tests. Educ Psychol Meas 1954; 14:3-19.
10. Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL (ed.), Educational Measurement. Washington, DC: American Council on education, 1971; 514-5.
11. Ebel RL. Essentials of Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall, 1972; 492-94.
12. Jones J, Hunter d. Consensus methods for medical and health services research. BMJ 1995; 311:376-80.

13. Clauser BE, Mee J, Baldwin SG, et al. Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: an experimental study. J EducMeas 2009; 46:390-407.
14. Hurtz GM, Auerbach MA. A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. Educ Psychol Meas 2003; 63:584–601.
15. Brandon PR. Conclusions about frequently studied modified Angoff standard setting topics. Appl Meas Educ 2004; 17:59–88.
16. Mee J, Clauser BE, Margolis MJ. The impact of process instructions on judges' use of examinee performance data in Angoff standard setting exercises. Educ Meas 2013; 32:27-35.
17. Margolis MJ, Clauser BE. The impact of examinee performance information on judges' cut scores in modified Angoff standard-setting exercises. Educ Meas 2014; 33:15-22.
18. Brennan RL, Lockwood RE. A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Appl Psychol Meas 1980; 4:219-40.
19. van der Linden WJ. A latent trait method for determining intrajudge inconsistency in Angoff and Nedelsky techniques of standard setting. J Educ Meas 1982; 19:295-308.
20. Chang L. Judgemental item analysis of the Nedelsky and Angoff standard-setting methods. Appl Meas Educ 1999; 12:151-65.
21. Ricker KL. Setting cut-scores: a critical review of the Angoff and modified Angoff methods. Alberta J Educ Res 2006; 52:53-64.
22. Clauser JC, Margolis MJ, Clauser BE. An examination of the repolcability of Angoff standard setting results within a generalizability theory framework. J Educ Meas 2014; 51:127-40.
23. Tannenbaum RJ, Kannan P. Consistency of Angoff-based standard-setting judgments: are item judgments and passing scores replicable across different panels of experts? Educ Assess 2015; 20:66-78.
24. Deunk MI, van Kuijk MF, Bosker RJ. The effect of small group discussion on cutoff scores during standard setting. Appl Meas Educ 2014; 27:77-97.
25. Raymond MR, Reid JB. Who made thee a judge? Selecting and training participants for standard setting. In Cizek GJ. Setting performance standards: Concepts, methods and perspectives. Mahwah NJ: Erlbaum, 2001; 119-57.
26. Smith RW, Davis-Becker SL, O'Leary LS. Combining the best of two standard setting methods: the ordered item booklet Angoff. J Appl Test Tech 2014; 15:18-26.
27. Cohen-Schotanus J, van der Vleuten CP. A standard setting method with the best performing students as point of reference: practical and affordable. Med Teach 2010; 32:154-60.