BMJ Open

# What should the standard be for passing and mastery on the Critical Thinking about Health Test? A consensus study

Allen Nsangi ,[1] Diana Aranza,[2] Roger Asimwe,[3,4] Susan Kyomuhendo Munaabi-Babigumira,[5] Judith Nantongo,[6] Lena Victoria Nordheim ,[7] Robert Ochieng,[8] Cyril Oyuga,[9] Innocent Uwimana,[10] Astrid Dahlgren,[11] Andrew Oxman[12]

**To cite:** Nsangi A, Aranza D, Asimwe R, *et al*. What should the standard be for passing and mastery on the Critical Thinking about Health Test? A consensus study. *BMJ Open* 2023;**13**:e066890. doi:10.1136/bmjopen-2022-066890

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-066890).

Received 03 August 2022
Accepted 10 February 2023



For numbered affiliations see end of article.

**Correspondence to**
Dr Andrew Oxman;
oxman@online.no

## ABSTRACT

**Objective** Most health literacy measures rely on subjective self-assessment. The Critical Thinking about Health Test is an objective measure that includes two multiple-choice questions (MCQs) for each of the nine Informed Health Choices Key Concepts included in the educational resources for secondary schools. The objective of this study was to determine cut-off scores for passing (the border between having and not having a basic understanding and the ability to apply the nine concepts) and mastery (the border between having mastered and not having mastered them).

**Design** Using a combination of two widely used methods: Angoff's and Nedelsky's, a panel judged the likelihood that an individual on the border of passing and another on the border of having mastered the concepts would answer each MCQ correctly. The cut-off scores were determined by summing up the probability of answering each MCQ correctly. Their independent assessments were summarised and discussed. A nominal group technique was used to reach a consensus.

**Setting** The study was conducted in secondary schools in East Africa.

**Participants** The panel included eight individuals with 5 or more years' experience in the following areas: evaluation of critical thinking interventions, curriculum development, teaching of lower secondary school and evidence-informed decision-making.

**Results** The panel agreed that for a passing score, students had to answer 9 of the 18 questions and for a mastery score, 14 out of 18 questions correctly.

**Conclusion** There was wide variation in the judgements made by individual panel members for many of the questions, but they quickly reached a consensus on the cut-off scores after discussions.

## STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ The cut-off scores were determined using a combination of robust methods.
⇒ The judging panel had content expertise and familiarity of context.
⇒ The panel included eight people, a number on the lower end of the spectrum recommended for both methods used.

## INTRODUCTION

Critical thinking is one of the most often included competencies in education systems the world over.[1–3] However, there is little agreement on its definition[4] or how it should be taught and evaluated.[5]

If health literacy is the ability to access, understand, appraise and apply health information,[6] then critical health literacy is potentially a higher order thinking process (critical thinking) that could be developed through education to critically appraise information relevant to health.[7]

Within the educational sector, critical thinking focuses on dispositions and abilities that help people to decide what to do or what to believe. While critical thinking and health are widely included in primary and secondary school curricula, critical health literacy or critical thinking about health is not.[8–10]

Individuals with higher levels of health literacy are more likely to make healthy choices in life. Poor health literacy has been found to be a barrier to access to basic health services such as screening,[8] lower adoption of preventive actions such as vaccination and insufficient understanding on the role of antibiotics.[11]

People with higher health literacy levels make better decisions when it comes to their health, they are more capable of adhering to treatments and they make more efficient use of resources.[9]

Health literacy assessment is recognised as an important consideration in delivering appropriately tailored effective healthcare and achieving better health outcomes.[12] However, health literacy assessment tools continue to primarily focus on individuals

and are slow in shifting from a medical perspective towards a societal one.[13]

The most frequently used tools reported in the literature are the Rapid Estimate of Adult Literacy in Medicine-Short Form, which tests reading ability through word recognition and pronunciation[14]; the Test of Functional Health Literacy in Adults, which requires patients to read and complete missing sections of selected passages of information to measure reading comprehension, as well as to read and apply the information on prescription labels and appointment slips to assess numeracy[15]; and the Newest Vital Sign, a quick assessment of reading comprehension and numeracy, requiring patients to read an ice cream nutritional label, then answer six problem-solving questions.[16 17]

All of these health literacy assessment tools, and other instruments used in children and adolescents[18] focus on functional literacy and do not assess critical health literacy particularly people's ability to appraise health information. Health literacy tools that include measures of critical health literacy, such as the European Health Literacy Survey Questionnaire, tend to rely on subjective self-assessment, which does not correlate with cognitive skills, rather than objective performance which does.[18]

The Informed Health Choices (IHC) project has developed learning resources based on a framework of concepts that people should understand and apply to assess healthcare claims and make informed health choices.[19] We initially developed resources for primary school children (10–12 years old). Those resources included a textbook, a workbook, a teachers' guide, a set of cards for one of the lessons and a classroom poster. The resources were found effective after being evaluated in a cluster randomised trial in Uganda.[20] Those resources addressed 12 IHC Key Concepts—concepts that students should understand and apply to assess healthcare claims and make informed health choices.[21 22]

Building on this body of work and context analyses in Kenya, Rwanda and Uganda,[23] we have developed digital resources for lower secondary school students (ages 14–16 years) in East Africa. Those resources, which address nine prioritised Key Concepts (table 1),[23] are being evaluated in cluster randomised trials.[24–26]

The primary outcome measure for the trials, an objective measure of critical health literacy, is a test with multiple-choice questions (MCQs) from the Claim Evaluation Tools item bank. The item bank contains MCQs that can be used to measure an individual's ability to apply each of the 49 IHC Key Concepts.[27] The MCQs can be used to assess learners' abilities, evaluate the effectiveness of interventions or map people's abilities.

The 'Critical Thinking about Health (CTH) Test' includes two MCQs for each of the nine Key Concepts addressed by the IHC lower secondary school resources. The primary outcome for the trials is the proportion of students who have a passing score on the CTH Test. Determining the proportion of students who pass requires determining a cut-off score, above which learners pass. In this context, a passing score indicates that learners:

- Have a basic understanding of the concepts and how to apply them.
- Do not need to repeat lessons or receive some other additional or alternative instruction.
- Are ready to go on to subsequent lessons that reinforce learning of the same concepts and introduce new concepts.

Setting a standard is essential to ensure that the test results will be meaningful, interpretable and defensible.[28]

**Table 1** Key Concepts included as learning goals in the IHC lower secondary school learning resources

| Higher-level concepts | Included Key Concepts |
|---|---|
| **Claims**<br>Claims about effects that are not supported by evidence from fair comparisons are not necessarily wrong, but there is an insufficient basis for believing them. | |
| Assumptions that treatments are safe or effective can be misleading. | 1. Do not assume that treatments are safe. |
| | 2. Do not assume that treatments have large, dramatic effects. |
| | 3. Do not assume that comparisons are not needed. |
| Trust based on the source of a claim alone can be misleading. | 4. Do not assume that personal experiences alone are sufficient. |
| Seemingly logical assumptions about treatments can be misleading. | 5. Do not assume that a treatment is better based on how new or technologically impressive it is. |
| | 6. Do not assume that a treatment is helpful or safe based on how widely used it is or has been. |
| **Comparisons**<br>To identify treatment effects, studies should make fair comparisons, designed to minimise the risk of systematic errors (biases) and random errors (the play of chance). | |
| Comparisons of treatments should be fair. | 7. Consider whether the people being compared were similar. |
| Descriptions of effects should reflect the risk of being misled by the play of chance. | 8. Be cautious of small studies. |
| **Choices**<br>What to do depends on judgements about a problem, the relevance of the available evidence, and the balance of expected benefits, harms and costs. | |
| Expected advantages should outweigh expected disadvantages. | 9. Weigh the benefits and savings against the harms and costs of acting or not. |

IHC, Informed Health Choices.

There is currently no relevant empirical literature on setting a standard for the CTH Test. Interpreting average differences in scores for a test or other continuous (or count) outcome measures is challenging.[29] It requires a basis for judging the importance of an average difference. For instance, 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 when comparing two groups of learners. The difference in the proportion of learners who have a passing score is more meaningful and easier to interpret than an average difference in test scores. However, one major statistical drawback for dichotomising this continuous variable may result in a loss of descriptive information on the performance of the study population. For example, the nature and extent of differences between individuals with poorer performance are lost when a cut-off score is dichotomised as having/not having passed with a passing or mastery score.[30]

### Objectives

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 secondary school resources.

## METHODS

We applied a modification of the Nedelsky's and Angoff's methods to determine an absolute standard.[31] Both methods rely on expert judges and the concept of individuals who are on the border of passing or failing. In the Nedelsky's method, judges eliminate response options that a borderline learner would be able to eliminate.[32] The chance of getting each question correct is then equal to one divided by the number of remaining response options. For example, if there are two remaining response options (one of which is the correct option), the chance of a borderline individual answering the question correct is one-half or 50%. The resulting cut-off score is then determined by adding up the probabilities for all the questions.

With Angoff's method, which is one of the most widely used, the judges assess the difficulty of each question as a whole.[33] The Angoff's method relies on subject matter experts who examine the content of each question (item) and then predict how many minimally qualified test takers would answer the item correctly.

Using a combination of Nedelsky's and Angoff's methods, starting with Nedelsky's method, the judges increased or decreased the probability of answering each question correctly based on an overall assessment. This gave them 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 may have made a question more or less difficult.

For each method, there are five stipulated steps:

1. Selection of judges.
2. Defining 'borderline' knowledge and ability.
3. Training of the selected judges in the use of the method.
4. Collection of their judgements.
5. Combining the judgements to determine a cut-off.

### Selection of the judges

In March 2022, we purposively selected and recruited four types of judges: lower secondary school teachers who participated in the pilot in each country to ensure that the judgements made were appropriate for the target audience and the context, health systems researchers and individuals who teach evidence-informed decision making, and curriculum developers and educational researchers with experience in evaluation of educational interventions designed to teach critical thinking skills (table 2).

The recommended number of judges when using the Angoff's method ranges from 5 to 30.[34] For this study, we recruited a total of eight judges,[35] a number that we considered to be manageable and adequate for making the required judgements while relying on our previous experience establishing a standard for passing and mastery for our earlier resources of primary school students.[36]

We had initially contacted nine individuals, all of whom agreed to participate in the process apart from one who

**Table 2** Judges

| | Sex | Country | Background |
|---|---|---|---|
| Judge 1 | M | Kenya | Curriculum specialist/educational researcher |
| Judge 2 | M | Rwanda | Curriculum specialist/educational researcher |
| Judge 3 | F | Norway | Educational researcher |
| Judge 4 | F | Norway/Uganda | Health systems researcher |
| Judge 5 | F | Uganda | Secondary school teacher |
| Judge 6 | M | Rwanda | Secondary school teacher |
| Judge 7 | F | Croatia | Health systems researcher |
| Judge 8 | M | Kenya | Secondary school teacher |

cited a busy work schedule, thereby leaving us with a number necessary to enable valid inferences to be made in addition to meaningful participation in the discussions.

A commonly held view in the scientific literature is that the resulting cut-off scores may be more accurate as the subject expertise of the judges increases, but that assertion has not been empirically confirmed.[35 37 38]

For this study, we aimed to ensure diversity within the panel of judges by selecting experienced individuals (5 or more years) with the following types of expertise:

- Health researchers and people who teach evidence-informed decision-making.
- Educational researchers with experience evaluating interventions to teach critical thinking skills.
- Curriculum or examination developers.
- Lower secondary school teachers in East Africa.

We invited at least one teacher from each country who participated in the pilot study of the learning resources to help ensure that the cut-offs are appropriate for the context in which the learning resources were to be evaluated. The context under consideration is lower secondary schools in East Africa, comprising of high teacher–student ratios of about 1:60 on average, limited resources and students with English as a second or third language. Judges were provided with instructions in advance on how judgements would be made (online supplemental appendix 1).

### Definition of borderline knowledge and ability

We defined a student on the border of passing as an individual who may or may not have a basic understanding of the concepts and the ability to apply them, may or may not need additional instruction, and may or may not be ready to go on to subsequent lessons. We defined a student on the border of master as an individual on the border between having mastered and not having mastered the nine key concepts, having a basic understanding of the concepts and how to apply them and having a clear understanding the concepts and how to apply them, and not needing and clearly not needing additional or alternative instruction and being ready to go on to other lessons which will reinforce learning of the same concepts and introduce new concepts.

We created personas that were characteristic of people on the border of passing and of people on the border of having mastered the concepts (online supplemental appendix 2).

### Training of the selected judges

The training of the selected judges occurred remotely, having sent the training materials (protocol, CTH Test, instructions and personas) a few days prior to the 1-hour online discussions where judges were given an opportunity to ask questions.

The instructions provided to the judges were discussed in detail before they started making their judgements (online supplemental appendix 1). The main objective of the training was to enable the judges to assess the difficulty of each question for two types of test takers: (1) ones who have a borderline understanding of the concepts they need to assess claims about treatment effects and (2) ones who have mastered the concepts. The judges took the CTH Test before they made judgements about the difficulty of the questions. On completion of the test, we did not assess their individual performance on the test but gave them the right answers to the questions as reference for when they made their judgements. We anticipated that giving them the correct answers after attempting the test themselves would help give them a sense of how difficult the questions were but individual assessments of their performance would not be necessary since some of the judges had participated in teaching the concepts in pilot schools thereby creating an unfair advantage.

The judges had a practice round with six MCQs with different degrees of difficulty before making their individual judgements. This exercise informed a discussion of what made a question easy or difficult. It also alerted them to their tendencies to be more or less pessimistic about the probability of a borderline student answering questions correctly in comparison with the other judges.

The ‘CTH Test’, although availed to the judging panel for purposes of setting a cut-off, is currently not available with this manuscript to avoid contamination pending the preplanned evaluations in the three East African countries for whose purpose a cut-off score is being set.

However, the Claim Evaluation Tools item bank found here is open access and free for non-commercial use.

### Collecting judgements and combining the judgements to determine a cut-off

The judges independently made their judgements for all 18 MCQs. One of us (AN) calculated the mean and the median for each MCQ and for the cut-off score. She presented these and the range to the judges. The judges were also shown the difficulty of the MCQs based on the results of the Rasch analysis after making their judgements (online supplemental appendix 3). AN and AO moderated an online discussion during which disagreements were discussed and resolved.

We used a nominal group technique to reach a consensus.[39] We initially shared all the judgements for each MCQ with the judges. We then invited those from each end of the spectrum to provide reasons for their judgements, before inviting others to comment. After the final cut-off score was agreed upon, we checked to make sure that all the judges agreed with the cut-off scores, and adjustments were made, if needed, based on the consensus of all the judges.

The same approach was used to determine a cut-off score for passing and for mastery.

### Patient and public involvement

There was no patient or public involvement in the study. In addition to participating in the process for the establishment of a standard for passing and mastery, study

participants have also been involved in the interpretation of the study results and the write-up of this manuscript.

## RESULTS

The discussions and consensus meetings were conducted online on 9 March and 22 March 2022. During the pilot, the judges agreed on the following:

- With a combination of prolonged school closures in East Africa due to the COVID-19 pandemic and English being a second or third language for many of the test takers, the judges agreed as a rule to always decrease the probability of answering a question correctly by at least 10% for both borderline and mastery test takers to account for reading errors.
- For purposes of determining the cut-off score, the judges agreed about the importance of keeping in mind the contexts in which we are using the test and the cut-off scores.

During our discussions about the judges' reasoning, we found that different judges had different reasons for their judgements, and each judge tended to apply the same reasoning across the MCQs.

Apart from one of the judges, who had participated in teaching the content to lower secondary school students during the piloting of an earlier version of the resources, the judges were not consistently biased towards underestimating or overestimating the difficulty of the MCQs. Although there were substantial differences in the panel's independent judgements about the difficulty of each MCQ (online supplemental appendix 4), there was less disagreement when the probabilities for each MCQ were summed up to determine the cut-offs, and the judges quickly came to a consensus about the difficulty of each MCQ and the cut-offs after a couple of deliberations, with each lasting at least an hour (table 3).

Following discussions of each MCQ and the cut-offs, the judges agreed that at least 9 questions out of 18 needed to be answered correctly to pass and at least 14 questions needed to be answered correctly to demonstrate mastery of the concepts.

## DISCUSSION

Empirical studies have shown that when judges use a common definition of minimally competent test takers, this tends to increase judgement consensus when determining cut-off scores.[38] Although there was substantial variation in the judges' independent assessments of the difficulty of each MCQ, the judges quickly reached a consensus, which is consistent with findings from multiple studies that determined cut-off scores.[38 40]

We provided the judges with performance data after they made their independent judgements, and made them aware that although the data provided an indication of the relative difficulty of the questions, it did not provide an indication of the probability of a borderline test taker answering a question correctly, since most of the Rasch analysis data[41] came from a mix of people, most of whom had not been taught and were not familiar with the IHC Key Concepts. Some studies have indicated that when judges view normative data, they tend to contaminate the process and systematically lower cut-off scores.[38 40] However, there is no indication that this occurred in this study.

When assembling a panel of judges, both the Angoff's and Nedelsky's methods recommend between 20 and 30 judges who are representative of the population to which the standards will be applied.[19] However, there is little agreement on the appropriate number of judges,[12 16] and several studies have found that between 5 and 10 judges is a manageable number and sufficient to determine cut-offs. In this study, the eight judges who participated in determining the cut-off scores came from different disciplines (education, health and evidence-based practice) and countries (Croatia, Kenya, Rwanda, Norway and Uganda).

Evidence suggests that cut-off judgements made using the Angoff's method are reproducible,[37] but there is a possibility for variability in cut-offs determined by different groups of judges as experience and context are brought into play.[42] There is no gold standard for setting a passing score. However, to ensure that the resulting cut-off is reproducible and unbiased, the approach that is used should ensure the credibility of the judges and use a systematic approach to collect their judgements. The key aspects to consider when selecting judges are their content expertise, familiarity with the context and examinees, and achieving a good balanced in gender and ethnicity.[43] This study met all these standards.

**Table 3** Individual and consensus summary judgements

| Judges | J-001 | J-002 | J-003 | J-004 | J-005 | J-006 | J-007 | J-008 | | Consensus |
|---|---|---|---|---|---|---|---|---|---|---|
| Rasch | | | | | | | | | 8 | |
| Chance | | | | | | | | | 6 | |
| Pass score (out of 18 questions) | 12 | 7 | 8 | 7 | 8 | 14 | 10 | 10 | | 8.76 (9) |
| Mastery score (out of 18 questions) | 14 | 14 | 13 | 11 | 13 | 17 | 16 | 11 | | 14.06 (14) |

Judges=J-001–J-008 (eight individuals).
Rasch=expected score based on difficulty of each question from Rasch analysis (proportion of participants who answered each question correctly).[41]
Chance=expected score by chance (guessing) alone.

### Strengths and limitations

We determined the cut-off scores using a combination of robust methods such as Angoff's and Nedelsky's while working with a panel of judges who had content expertise and were familiar with the context. The cut-offs were established for students in lower secondary schools in East Africa. It is uncertain whether the same cut-offs are appropriate for other contexts.

However, the methods used in this study are robust and efficient and could be used in other settings, as well as for other tests using questions from the Claim Evaluation Tools database or other MCQs.[36]

Although the number of judges recommended by both the Angoff's and Nedelsky's methods ranges from 5 to 30, we were on the lower end of the spectrum with only eight members on the judging panel, a number we found manageable but may have left out significant contributions from others to the judgements.

## CONCLUSION

Although there was wide variation in many of the individual judgements, it was possible to reach a consensus on the cut-off scores for passing and mastery in an online meeting that lasted less than 90 min.

The use of a combination of the Angoff's and Nedelsky's methods, in addition to initial agreement on some general guidance following a pilot, ensured an appropriate process that resulted in absolute standards for having a basic understanding (passing) and mastery of the nine concepts addressed in the IHC secondary school resources.

**Author affiliations**
[1]Department of Medicine, Makerere University College of Health Sciences, Kampala, Uganda
[2]University Department for Health Studies, University of Split, Split, Croatia
[3]Lower Secondary School Section, Group Scolaire Nduba, Kigali, Rwanda
[4]Secondary School Teaching, Ministry of Education, Kigali, Rwanda
[5]Department of Global Health, Norwegian Institute of Public Health, Oslo, Norway
[6]Biology Department, Baptist High School, Kitebi, Uganda
[7]Department of Health and Functioning, Faculty of Health and Social Sciences, Western Norway University of Applied Sciences, Bergen, Norway
[8]Lower Secondary Section, Kibos Secondary School, Kondele, Kenya
[9]Research and Knowledge Management Department, Kenya Institute of Curriculum Development, Nairobi, Kenya
[10]Basic Education, Rwanda Education Board, Kigali, Rwanda
[11]Faculty of Health Sciences, Oslo Metropolitan University, Oslo, Norway
[12]Centre for Informed Health Choices, Norwegian Institute of Public Health, Oslo, Norway

**Twitter** Allen Nsangi @AllenNsangi and Roger Asimwe @AsiimweRoger10

**Acknowledgements** We are greatly indebted to the Informed Health Choices (IHC) Team for their invaluable feedback during the planning of this study. In addition, we would like to thank the students and teachers who participated in the pilot and validation (Rasch analysis study) of the Critical Thinking about Health Test whose findings we drew upon to provide context for the judges on the panel.

**Contributors** AN, AO and AD were responsible for study conception, wrote the protocol, conducted the study, and led data acquisition, analysis and interpretation. DA, RA, SKM-B, JN, LVN, RO, CO and IU provided feedback during the process, and participated in data acquisition and interpretation. AN drafted this paper, while AO, AD, DA, RA, SKM-B, JN, LVN, RO, CO and IU provided substantial input to the draft. AN is the article guarantor.

**Funding** This study was funded by the Norwegian Research Council (project number: 284683, grant number 69006) awarded to AO through the Norwegian Institute of Public Health, in collaboration with Makerere University, Uganda, Tropical Institute of Community Health and Development, Kenya and the University of Rwanda, Rwanda.

**Disclaimer** The funder had no role in the study design, preparation of the manuscript and publication decision.

**Competing interests** None declared.

**Patient and public involvement** No patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

**Patient consent for publication** Not required.

**Ethics approval** We obtained ethics approval from the following institutions: (1) Rwanda National Ethics Committee (approval number 916/RNEC/2010) for the Rwandan study site; (2) Masinde University of Science and Technology Institutional Ethics Review Committee and the Kenyan National Commission for Science, Technology and Innovation (approval number NACOSTI/P119/1986) for the Kenyan study site; (3) Makerere University School of Medicine Research Ethics Committee and the Uganda National Council of Science and Technology (reference number HS91ES) for Uganda. Participants gave informed consent to participate in the study before taking part.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available upon reasonable request. Data have been made available as an appendix to this manuscript. Any extra data will be made available on reasonable request.



**ORCID iDs**
Allen Nsangi http://orcid.org/0000-0001-8702-9217
Lena Victoria Nordheim http://orcid.org/0000-0001-7370-1121

## REFERENCES

1 Erstad O, Voogt J. The twenty-first century curriculum: issues and challenges. In: Voogt J, Knezek G, Christensen R, et al., eds. *Second Handbook of Information Technology in Primary and Secondary Education* Cham, Switzerland: Springer International Publishing, n.d.: 2018. 19–36.
2 Care E, Anderson K, Kim H. Visualizing the breadth of skills movement across education systems. Washington, DC Brookings Institution; 2016.
3 Voogt J, Roblin NP. A comparative analysis of international frameworks for 21st century competences: implications for national curriculum policies. *Journal of Curriculum Studies* 2012;44:299–321.
4 Geng F. An content analysis of the definition of critical thinking. *ASS* 2014;10:19.
5 Abrami PC, Bernard RM, Borokhovski E, *et al*. Strategies for teaching students to think critically. *Review of Educational Research* 2015;85:275–314.
6 McLaughlin M, DeVoogd G. Critical literacy as comprehension: expanding reader response. *J Adolesc Adult Lit* 2004;48:52–62.
7 Sykes S, Wills J, Rowlands G, *et al*. Understanding critical health literacy: a concept analysis. *BMC Public Health* 2013;13:150.

8 Kobayashi LC, Wardle J, von Wagner C. Limited health literacy is a barrier to colorectal cancer screening in england: evidence from the english longitudinal study of ageing. *Prev Med* 2014;61:100–5.
9 Santos P, Sá L, Couto L, *et al*. Health literacy as a key for effective preventive medicine. *Cogent Social Sciences* 2017;3:1407522.
10 Miller TA. Health literacy and adherence to medical treatment in chronic and acute illness: a meta-analysis. *Patient Educ Couns* 2016;99:1079–86.
11 Berkman ND, Sheridan SL, Donahue KE, *et al*. Low health literacy and health outcomes: an updated systematic review. *Ann Intern Med* 2011;155:97–107.
12 Guzys D, Kenny A, Dickson-Swift V, *et al*. A critical review of population health literacy assessment. *BMC Public Health* 2015;15:215.
13 Freedman DA, Bess KD, Tucker HA, *et al*. Public health literacy defined. *Am J Prev Med* 2009;36:446–51.
14 Davis TC, Long SW, Jackson RH, *et al*. Rapid estimate of adult literacy in medicine: a shortened screening instrument. *Fam Med* 1993;25:391–5.
15 Parker RM, Baker DW, Williams MV, *et al*. The test of functional health literacy in adults. *J Gen Intern Med* 1995;10:537–41.
16 Osborne RH, Batterham RW, Elsworth GR, *et al*. The grounded psychometric development and initial validation of the health literacy questionnaire (HLQ). *BMC Public Health* 2013;13:658.
17 Moore V. Assessing health literacy. *The Journal for Nurse Practitioners* 2012;8:243–4.
18 Guo S, Armstrong R, Waters E, *et al*. Quality of health literacy instruments used in children and adolescents: a systematic review. *BMJ Open* 2018;8:e020080.
19 Chalmers I, Oxman AD, Austvoll-Dahlgren A, *et al*. Key concepts for informed health choices: a framework for helping people learn how to assess treatment claims and make informed choices. *BMJ Evid Based Med* 2018;23:29–33.
20 Nsangi A, Semakula D, Oxman AD, *et al*. Effects of the informed health choices primary school intervention on the ability of children in uganda to assess the reliability of claims about treatment effects: a cluster-randomised controlled trial. *Lancet* 2017;390:374–88.
21 Oxman AD, Chalmers I, Austvoll-Dahlgren A, *et al*. Key concepts for assessing claims about treatment effects and making well-informed treatment choices. *F1000Res* 2018;7:1784.
22 Oxman AD, Chalmers I, Austvoll-Dahlgren A, *et al*. Key concepts for assessing claims about treatment effects and making well-informed treatment choices [version 2]. *F1000Res* 2018;7.
23 Agaba JJ, Chesire F, Michael M, *et al*. Prioritisation of informed health choices (IHC) key concepts to be included in lower-secondary school resources: a consensus study. *Public and Global Health* [Preprint].
24 Chesire F, Kaseje M, Ochieng M, *et al*. Effects of the informed health choices secondary school intervention on the ability of lower secondary students in kenya to think critically about health information and choices: protocol for a cluster-randomized trial. *Zenodo* 2022.
25 Mugisha M, Nyirazinyoye L, Simbi CMC, *et al*. Effects of using the informed health choices digital secondary school resources on the ability of rwandan students to think critically about health: protocol for a cluster-randomised trial. *Zenodo* 2022.
26 Ssenyonga R, Sewankambo NK, Mugagga SK, *et al*. Does the use of the informed health choices teaching resources improve the secondary students' ability to critically think about health in uganda? A cluster randomised trial protocol. *Zenodo* 2022.
27 Austvoll-Dahlgren A, Guttersrud Ø, Nsangi A, *et al*. Measuring ability to assess claims about treatment effects: a latent trait analysis of items from the " claim evaluation tools " database using Rasch modelling. *BMJ Open* 2017;7:e013185.
28 Lane AS, Roberts C, Khanna P. Do we know who the person with the borderline score is, in standard-setting and decision-making. *Health Prof Educ* 2020;6:617–25.
29 Guyatt GH, Thorlund K, Oxman AD, *et al*. GRADE guidelines: 13. preparing summary of findings tables and evidence profiles-continuous outcomes. *J Clin Epidemiol* 2013;66:173–83.
30 Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ* 2006;332:1080.
31 Livingston SA, Zieky MJ. *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service, 1982.
32 Nedelsky L. Absolute grading standards for objective tests. *Educational and Psychological Measurement* 1954;14:3–19.
33 Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL, ed. *Educational Measurement*. Washington, DC: American Council on education, 1971: 514–5.
34 Shulruf B, Wilkinson T, Weller J, *et al*. Insights into the angoff method: results from a simulation study. *BMC Med Educ* 2016;16:134.
35 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 Theory Pract* 2008;13:11–24.
36 Davies A, Gerrity M, Nordheim L, *et al*. Measuring ability to assess claims about treatment effects: establishment of a standard for passing and mastery. *IHC Working Paper* 2017.
37 Clauser JC, Margolis MJ, Clauser BE. An examination of the replicability of angoff standard setting results within a generalizability theory framework. *Journal of Educational Measurement* 2014;51:127–40.
38 Hurtz GM, Auerbach MA. A meta-analysis of the effects of modifications to the angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement* 2003;63:584–601.
39 Jones J, Hunter D. Consensus methods for medical and health services research. *BMJ* 1995;311:376–80.
40 Zieky M, Perie M. A primer on setting cut scores on tests of educational achievement. 2006.
41 Dahlgren A, Semakula D, Chesire F, *et al*. Critical thinking about treatment effects in eastern africa: development and evaluation of an assessment tool using rasch analysis. *Plos One* 2022.
42 Tannenbaum RJ, Kannan P. Consistency of angoff-based standard-setting judgments: are item judgments and passing scores replicable across different panels of experts? *Educational Assessment* 2015;20:66–78.
43 Downing SM, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teach Learn Med* 2006;18:50–7.