







## RESEARCH ARTICLE

# Critical thinking about treatment effects in Eastern Africa: development and Rasch analysis of an assessment tool

## [version 1; peer review: awaiting peer review]

Astrid Dahlgren <sup>1,2</sup>, Daniel Semakula <sup>3</sup>, Faith Chesire<sup>4</sup>, Michael Mugisha<sup>5</sup>, Esther Nakyejwe<sup>3</sup>, Allen Nsangi <sup>3</sup>, Laetitia Nyirazinyoye<sup>5</sup>, Marlyn A. Ochieng <sup>4</sup>, Andrew David Oxman<sup>1</sup>, Ronald Ssenyonga<sup>3</sup>, Clarisse Marie Claudine Simbi<sup>5</sup>

<sup>1</sup>Faculty of Health Sciences, Oslo Metropolitan University, Oslo, Norway

<sup>2</sup>Norwegian Institute of Public Health, Oslo, Norway

<sup>3</sup>Department of Medicine, College of Health Sciences, Makerere University, Kampala, Uganda

<sup>4</sup>Tropical Institute of Community Health and Development, Kisumu, Kenya

<sup>5</sup>School of Public Health, College of Medicine and Health Sciences, University of Rwanda, Kigali, Rwanda

---

**v1** First published: 26 Jul 2023, 12:887  
<https://doi.org/10.12688/f1000research.132052.1>  
Latest published: 26 Jul 2023, 12:887  
<https://doi.org/10.12688/f1000research.132052.1>

---

### Abstract

**Background:** Every day we are faced with different treatment claims, in the news, in social media, and by our family and friends. Some of these claims are true, but many are unsubstantiated. Without being supported by reliable evidence such guidance can lead to waste and harmful health choices. The Informed Health Choices (IHC) Network facilitates development of interventions for teaching children and adults the ability to assess treatment claims ([informedhealthchoices.org](http://informedhealthchoices.org)). Our objective was to develop and evaluate a new assessment tool developed from the item bank for use in an upcoming trial of lower secondary school resources in Uganda, Kenya, and Rwanda.

**Methods:** A cross-sectional study evaluating a questionnaire including two item-sets was used. The first evaluated ability using multiple-choice questions (scored dichotomously) and the other evaluated intended behaviour and self-efficacy (measured using Likert scales). This study was conducted in Uganda, Kenya, and Rwanda in 2021. We recruited children (over 12 years old) and adults through schools and our networks. We entered 1,671 responses into our analysis. Summary and individual fit to the Rasch model (including Cronbach's Alpha) were assessed using the RUMM2030 software.

**Results:** Both item-sets were found to have good fit to the Rasch model and were acceptable to our target audience. The reliability was good (Cronbach's alpha >0.7). Observations of the individual item and person fit provided us with guidance on how we could improve the design, scoring, and administration of the two item-sets. There was no

### Open Peer Review

**Approval Status** *AWAITING PEER REVIEW*

Any reports and responses or comments on the article can be found at the end of the article.

local dependency in either of the item-sets, and both item-sets were found to have acceptable unidimensionality.

**Conclusion:** To our knowledge, this is the first instrument validated for measuring ability to assess treatment claims in Uganda, Kenya and Rwanda. Overall, the two item-sets were found to have satisfactory measurement properties.

### Keywords

health literacy, Rasch analysis, critical thinking, informed choice, evidence-based practice



This article is included in the **Global Public Health gateway**.

**Corresponding author:** Astrid Dahlgren ([astridad@gmail.com](mailto:astridad@gmail.com))

**Author roles:** **Dahlgren A:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Semakula D:** Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing; **Chesire F:** Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing; **Mugisha M:** Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing; **Nakyejwe E:** Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing; **Nsangi A:** Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing; **Nyirazinyoye L:** Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing; **Ochieng MA:** Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing; **Oxman AD:** Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Writing – Review & Editing; **Ssenyonga R:** Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing; **Simbi CMC:** Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This study was funded by the Research Council of Norway (Project number 284683, grant number 69006). *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2023 Dahlgren A *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Dahlgren A, Semakula D, Chesire F *et al.* **Critical thinking about treatment effects in Eastern Africa: development and Rasch analysis of an assessment tool [version 1; peer review: awaiting peer review]** F1000Research 2023, 12:887 <https://doi.org/10.12688/f1000research.132052.1>

**First published:** 26 Jul 2023, 12:887 <https://doi.org/10.12688/f1000research.132052.1>

## Introduction

Every day we are faced with different treatment claims, in the news, in social media, and by our family and friends. Some of these claims are true, but many are unsubstantiated.<sup>1,2</sup> Without being supported by reliable evidence such guidance can lead to waste and harmful health choices.<sup>3,4</sup> Thus, improving people's ability to assess whether treatment claims are based on reliable evidence may lead to better health outcomes. The spread of misinformation during the Covid-19 pandemic has further emphasized the importance of promoting critical thinking and science literacy as a public health initiative.<sup>5,6</sup>

The Informed Health Choices (IHC) Network facilitates development of interventions for teaching children and adults the ability to assess treatment claims ([informedhealthchoices.org](https://informedhealthchoices.org)). We have developed a list of Key Concepts that people need to know to be able to assess claims about treatment effects.<sup>7</sup> By 'treatment' we refer to any intervention (action) intended to improve health, including preventive, therapeutic, and rehabilitative interventions, and public health or health system interventions. In two recent randomized trials in Uganda, we found that primary school children and their parents could be taught to apply these concepts.<sup>8,9</sup> Currently we are preparing for a new trial in Kenya, Rwanda, and Uganda to evaluate a set of educational resources for lower secondary schools (the IHC secondary school resources).

The Claim Evaluation Tools item bank was first developed for use in the abovementioned trials in Uganda, evaluating learning outcomes in primary school children and their parents.<sup>8,9</sup> We also developed the item bank so that it could be used as a flexible resource for teachers and researchers, enabling them to design their own instrument for their own purposes.<sup>10,11</sup> The item bank can be used for creating tests in schools (including higher education) and for research purposes in, for example, surveys and randomized trials.

Since it was first developed, the item bank has been periodically revised to reflect changes we have made to the Key Concepts list. Since our first trials in Uganda, researchers have developed instruments using items from the item bank in other contexts, including China, Mexico, and Norway.<sup>12-14</sup> Other studies are underway in Croatia and the USA. Currently, the item bank includes more than 200 items, with three to four multiple-choice questions (MCQs) available for assessing knowledge and the ability to apply each concept in the list. The item bank also includes a sample of literacy questions for use in contexts where reading ability may be a barrier for responding to the MCQs. It also includes items for assessing people's intended behaviours and self-efficacy (scored on 5-point Likert scales). All items are written in plain language and are suitable for both children and adults.

In the present study, our objective was to develop and evaluate the psychometric properties of a new assessment tool developed from the item bank for use in Uganda, Kenya, and Rwanda. This outcome measure will be used in randomised trials of the IHC lower secondary school resources.

## Methods

Below we describe how we designed the questionnaire, how it was administered, and how we analysed and report the data. The protocol and underlying data for this study has been published.<sup>34,35</sup>

### Designing the questionnaire

For this study we included both ability items and the items measuring intended behaviour and self-efficacy.

We planned on removing MCQs with sub-optimal measurement properties based on the results of this study. Therefore, we included more MCQs than we plan to use in the trial (two MCQs per Key Concept). The educational intervention we will evaluate in the randomised trials addresses nine Key Concepts (Table 1). For each of those concepts, we included three MCQs in the questionnaire, a total of 27 MCQs assessing ability. All MCQs included 3 response options.

We included three items that assess intended behaviour and four items that assess self-efficacy. The Likert scales include four response options ranging from very likely to very unlikely (intended behaviour) or very difficult to very easy (self-efficacy), and a fifth option: 'I don't know'.

In addition, we included demographic questions asking about gender, age, educational level, country of residence, training in research methods, and experience with participation in randomised trials. Gender, age, and country of residence were important for the psychometric analysis (testing for differential item functioning). The other background factors were used to ascertain that we were able to recruit people with a spread in ability level (ability to assess treatment claims). Level of education and familiarity with research methods have been shown to be associated with more correct answers.<sup>14</sup>

In preparation for this study, we conducted cognitive interviews and piloted the questionnaire with individuals from our potential target groups in Uganda, Kenya, and Rwanda.<sup>11,15</sup> The objective was to get feedback from members of our

**Table 1. Key Concepts included as learning goals in the Informed Health Choices (IHC) lower secondary school learning resources.**

Higher level concepts	Included Key Concepts
<b>Claims</b> 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.	<ol style="list-style-type: none"> <li>1. Do not assume that treatments are safe.</li> <li>2. Do not assume that treatments have large, dramatic effects.</li> <li>3. Do not assume that comparisons are not needed.</li> </ol>
Trust based on the source of a claim alone can be misleading.	<ol style="list-style-type: none"> <li>4. Do not assume that personal experiences alone are sufficient.</li> </ol>
Seemingly logical assumptions about treatments can be misleading.	<ol style="list-style-type: none"> <li>5. Do not assume that a treatment is better based on how new or technologically impressive it is.</li> <li>6. Do not assume that a treatment is helpful or safe based on how widely used it is or has been.</li> </ol>
<b>Comparisons</b> To identify treatment effects, studies should make fair comparisons, designed to minimize the risk of systematic errors (biases) and random errors (the play of chance).	
Comparisons of treatments should be fair.	<ol style="list-style-type: none"> <li>7. Consider whether the people being compared were similar.</li> </ol>
Descriptions of effects should reflect the risk of being misled by the play of chance.	<ol style="list-style-type: none"> <li>8. Be cautious of small studies.</li> </ol>
<b>Choices</b> 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.	<ol style="list-style-type: none"> <li>9. Weigh the benefits and savings against the harms and costs of acting or not.</li> </ol>

target groups in the three contexts on the acceptability and relevance of the terminology and formats used in the questionnaire. Even though the items included in the Claim Evaluation Tools item bank have previously gone through an extensive development process in Uganda, we considered it important to get feedback from people in our target groups in Rwanda and Kenya, where the items had not been tested before.

We recruited schools in May- August 2021 through the project’s teacher networks. In the interviews the students were encouraged to think aloud about how they understood the scenarios and response options, and to identify any issues they had regarding comprehension of terminology or format. The researcher noted down all identified issues. All feedback was summarised by the lead investigators and the findings was discussed in the project group including the research teams in all three contexts.

Piloting took place in a classroom setting. The purpose and instructions of the test was introduced to the students by a member of the research team in collaboration with the teacher, observations were made regarding time taken to complete the questionnaire and comprehension of the format (incorrectly filled in response options).

Findings coming out of the interviews and pilots led to only minor changes, such as changing some of the names and other terminology used in the MCQs to improve familiarity in the two new contexts. We also changed the format of the intended behaviour and self-efficacy items from a traditional Likert-scale to resemble a multiple-choice format, keeping the same response options (Figure 1).

We made that change because the Likert-scale format was unfamiliar to some of the students in the three contexts, and the MCQ format was more familiar and acceptable to the students. The pilot studies also provided us with information about the time needed to complete the questionnaire (between 30 and 60 minutes) and what we could expect in terms of missing responses in the upcoming trial.

Think about a sickness that you might get. Imagine someone claiming (saying) that a treatment might help you get better.

**4.5 Question:** How likely are you to **find out what the claim was based on** (for example by asking the person making the claim)?

*Options:*

- A) Very unlikely
- B) Unlikely
- C) Likely
- D) Very likely
- E) I don't know

**Figure 1. Example of an intended behaviour item.**

Previously, several tests have been developed from the claim evaluation tools item bank. The test developed for this study was named the Critical Thinking about Health test. A copy of the test evaluated in this study is available as extended data.<sup>36</sup>

#### Inclusion criteria

There is no gold standard for the number of respondents needed for Rasch analysis. This is a pragmatic judgement considering the number of items evaluated and the statistical power needed to identify item bias resulting from background variables.<sup>16–18</sup> Rasch analysis does not require a representative sample. However, the sample should include enough people to allow for evaluating differential functioning and a spread in ability. Studies have found that a sample of 200-250 people per group is suitable for detecting differential item functioning (DIF).<sup>19,20</sup> We expected both item-sets to work in the same way for children and adults and to have no differential functioning by gender.<sup>11</sup> For this evaluation, we also needed a sample of people with different ability to assess treatment claims. There are few background variables that may predict ability to assess treatment claims, but higher education involving training in statistics or research methods may be a factor.<sup>14</sup> Consequently, we estimated that recruiting approximately 500 people in each country, with an equal distribution of men and women, and lower secondary school students and adults would be adequate (Table 2). We also made sure to recruit people from higher education contexts, through the university networks in each context, as well as people in our local communities, social media, and students from schools participating in piloting of the educational intervention. We commenced data collection in July 2021 and was completed December in the same year.

#### Recruitment

All recruitment and data collection were done during lock down due to COVID-19, leading us to use varied strategies for recruiting our respondents.

In Uganda we recruited participants using our networks there including teachers, students, and National advisory panel networks. For students, we used three strategies, including visiting students at their homes, reaching out through the student network, and also requested teachers who were conducting online revision classes to introduce us to their students via the platforms to introduce the project and share the questionnaire link via WhatsApp or Telegram (both media apps for communication) after obtaining consent. For adults, we recruited people with higher education qualifications through

**Table 2. Overview of sample to be recruited.**

Participants	Kenya	Rwanda	Uganda
Secondary school children	>250	>250	>250
Adults	>250	>250	>250

university platforms i.e., the University faculty platforms, a PhD forum which has over 40 PhD fellows, students studying medicine WhatsApp groups, and a teachers' network WhatsApp group. However, for the local communities, we visited food and clothes markets and asked them to complete the questionnaires. All data collection was done in the central region (Kampala and Wakiso) and the northern region (in Gulu district) of Uganda.

In Kenya we recruited students from three schools that participated in piloting the IHC secondary school resources. In those schools, we purposively included all the participants from one stream except those that had been selected for the pilot. Each school had about three-four classes and each class had about 40 students. For adults, we included the student's institution of tertiary education and members of the community with low education levels (secondary and below), and those that could read and owned a Smartphone. For the students, we purposively included students from two faculties (Health and Arts and Sciences). Through the Dean of students, we invited them to a meeting where we introduced the project, outcome measure and sought their verbal consent. We then shared the link to the test and asked them to log in and participate. For community members, we used our database to recruit members that were actively involved in the institute's previous and ongoing community-based projects in rural settings in Butere sub-County. Although we reached out to many members, only a few members responded thus we resorted to recruit more from the student's fraternity (pursuing diploma and certificate courses). We used a similar recruitment and consenting process described for the students above.

In Rwanda, for adults, we used WhatsApp and recruited using the snowballing method through our networks, including the projects teachers' network and students' network in Rwanda. The teachers network included lower secondary school teachers who were from different schools, and they varied in terms of work experience, age, subject area and schools they teach from. Similarly, the students' network included students from similar schools as members of teacher's network. They also varied in their age, sex, and history of school performance (high or low performing students). We also used emails and reached out to adults who work or previously worked with the school of public health researchers in Rwanda. We also engaged a teacher's network who also responded to the test. We recruited students through schools that participated in the development and pilot of the intervention in Kigali city and surrounding neighborhoods.

### Data collection

Most of the data collection was done online, using a service hosted by the University of Oslo (Nettskjema). One small sample (students in Kenya) used paper questionnaires in a classroom setting and administrated as an exam as part of pilot testing of the IHC secondary school resources. The test was administrated by a teacher under the instructions of the research team. The paper questionnaires were scanned and added to the data collected online.

### Ethical statement

Ethical approval was obtained from the relevant authorities in each country; Masinde Muliro University of Science and Technology, Institutional Ethics Review Committee (MMUST/IERC/75/19, License No: NACOSTI/P/21/8103) the Rwanda National Ethics Committee 916/RNEC/2019, School of Medicine Research Ethics Committee (REC REF 2020-139)/Uganda National Council of Science and Technology (HS916ES).

All participants were given written information about the purpose of the study and that participation was voluntary, and how the findings would be used to improve the validity and reliability of the Critical Thinking about Health test. Children participating through their schools were also given oral information. We obtained written consent from all adult participants, the minor's guardians, and written assent from the minors.

Since this was a knowledge test, just as a regular school exam, this study did not collect any personal or other sensitive information that could lead to identification of the respondents. None of the members of this project group had access to information that could identify individual participants during or after data collection.

### Rasch analysis

Rasch analysis is a dynamic way of developing measurement tools with construct validity.<sup>14</sup> The approach is used to address important measurement issues required for validating an outcome measure, including internal construct validity (by testing for unidimensionality), invariance of the items (item-person-interaction), and item bias (differential item function).<sup>21,22</sup>

We imported the data from Excel (version 2208) into RUMM2030 (<https://www.rummlab.com.au/>) and followed the basic steps of Rasch analysis as recommended in the literature.<sup>21,23</sup> R is a freely accessible software environment for statistical computing and graphics including Rasch analysis that can be used to run a similar analysis (<https://www.r-project.org/>). We analysed the two item-sets separately based on the assumption that these measure different underlying

traits. The MCQs were scored dichotomously as correct or incorrect. We applied the polytomous model to the intended behaviour and self-efficacy items.<sup>22</sup> When entered into RUMM2030, missing data was coded as “0”.

The first step in the analysis involved exploring the class interval structure (number and size of ability groups) and the summary statistics (person-Item distribution). In Rasch analysis, the ratio between any two items should be constant across different ‘ability’ groups. The response patterns to an item-set is tested against what is expected by the model which is a probabilistic form of Guttman scaling.<sup>21</sup> In other words, the easier the item is, the more likely it will be ‘passed’, and the more able the person is the more likely he or she will pass.<sup>21</sup> We explored this relationship using the summary statistics function in RUMM2030.<sup>23</sup> In RUMM2030, the item-person interaction is presented on a logit scale, where the mean item location is ‘0’. If the instrument is a well-targeted measure (not too easy or too difficult), the mean location for individuals would be around the value of zero.<sup>22</sup> If the person location is higher than zero, this indicates that the test is easy, if the person location is lower than zero this indicates that the test is difficult. The item and person fit residual statistics assess the degree of divergence (or residual) between the expected and observed data for each person item when summed for all items and all individuals respectively for each test set.<sup>22</sup> In RUMM2030 this is reported as an approximate z-score, representing a standardized normal distribution.<sup>22</sup> Ideally, item and person fit should have a mean of zero and a standard deviation of one.<sup>22</sup>

We calculated Cronbach’s alpha to assess the reliability of both item-sets by removing missing data. A Cronbach’s alpha above 0.7 was considered acceptable.<sup>22</sup>

The principal component analysis/t-test protocol is used to test the hypothesis of unidimensionality. This is done by identifying the two most divergent item subsets (using the residual principal component function in RUMM2030), and then calculating t-tests.<sup>22</sup> If  $\leq 5\%$  of tests are significant, strict unidimensionality can be inferred.<sup>24</sup> However, the concept of ‘unidimensionality’ is not ‘definite’ but relative and should be supplemented with quantitative or qualitative interpretation of the explicit variable definition and considering the context and purpose of the measurement.<sup>24,25</sup>

We tested for local dependency by using the residual correlations function in RUMM2030. Data from this output was copied into Excel (version 2208) and any residual correlations greater than 0.2 above the average was considered as potential problematic dependency.<sup>22</sup>

We identified individuals and items with ‘misfit’ to the Rasch model by chi-square statistics and by exploring the fit residuals. Items with statistically significant chi-square probabilities do not fit the model at 0.01 significance level, items within a  $\pm 2.5$  fit residual range are considered to be potentially problematic.<sup>22</sup> Similarly, individuals with a fit residual of  $\pm 2.5$  were considered as not fitting the model. Such extreme values can be an indication of, for example, guessing or copying, and that the item-set is not appropriate.

We examined differential item functioning (DIF) by age, gender, and country of residence. It was our objective to include only items that could be applied fairly across these demographic variables. Ideally, all items in the Claim Evaluation Tools item bank are expected to work in the same way for men and women, and across age groups. There are two types of DIF. Uniform DIF is when the difference between groups for an item is systematic - for example adults having systematically higher ability compared to lower secondary school students. This is less problematic (when it is known) than non-uniform DIF, where the difference between groups on an item is inconsistent across ability groups.<sup>21</sup> For this study, we considered non-uniform DIF as unacceptable. We predicted that we would find uniform DIF by country, as we know from other studies that there are differences in ability-by-concept across countries.<sup>14</sup> Uniform DIF by gender and age was unwanted but would be considered in relation to the other findings from the Rasch analysis. The reason for this was that the questionnaire will be used for measuring differences between an intervention and a comparison group, and systematic DIF would therefore not be a problem in our study.

In the item characteristic curve plot the expected scores and the observed scores for the class intervals of the different ability levels are displayed. We observed the item characteristic curve for each item and made note of items that showed under-discrimination, over-discrimination, or had several deviating ability groups.<sup>22</sup> We considered items with under-discrimination and classic over-discrimination for removal. Marginal over-discrimination was not considered to be a problem for our purposes.

For the polytomous items we explored the threshold ordering (fit to the expected logical order of the response options) to check for disordered thresholds. Disordered thresholds suggest that the scoring categories are not progressing as expected, and that the item is not working properly.<sup>22</sup>

This study follows the STROBE-reporting standards.<sup>38</sup>

## Results

### Summary statistics

A total of 1,671 responses were entered into the analysis distributed across 10 ability groups identified by the RUMM2030 software of which 49% were women and 40% were young people (under 18). Of these, 35% were from Kenya, 34% from Uganda, and 31% from Rwanda. Missing data was minimal only 0.004%, and thus had no impact on the analysis.

The person-item distribution shows that both item sets are well targeted (mean person location was -0.218 for the ability item set and 0.084 for the Likert item-set).

For the ability items, the person fit residual was -0.204 (SD 0.741) and thus showed satisfactory fit to the model. The items' fit residual was 0.712 (SD 2.235) and warranted further investigation in subsequent analyses.

For the Likert items, the item fit residual was 0.543 (SD 0.938), indicating reasonable fit. However, the high standard deviation for the person fit residual (-0.546, SD 1.783) suggested some misfit to the model.

Both item-sets were found to be reliable, with a Cronbach's alpha of 0.72 and 0.79 for the ability and Likert item-sets respectively.

### Individual person and item fit

In the analysis of the ability item-set, we identified one person with a highly negative fit residual (adult, female, Rwanda) and two with highly positive fit residuals (male, young person, Rwanda and adult, female, Rwanda). Of the 27 MCQs, three items had extreme negative values, and four items had extreme positive values.

There were no items with extreme values in the Likert item-set. However, several misfitting persons were identified (296 individuals) with high negative residuals and two individuals with high positive residuals.

### DIF and item characteristic curve-analyses

The majority of the ability items had a good fit to the item characteristics curve (Figure 2). Four items showed evidence of classic overdiscrimination, of which two of these also had very high negative fit residuals (Figure 3). Four items showed sign of classic underdiscrimination and were considered candidates for removal (Figure 4). Most Likert-items showed a good fit, although two items were slightly overdiscriminating, this was considered acceptable.

In the DIF analysis of the 27 ability items, two items showed uniform DIF by gender (one item where males did systematically better and one where females had higher ability). Three items showed DIF by age, of which two were uniform (one item where young people performed better and one item where adults had higher ability). One item had non-uniform DIF by age. Uniform DIF by country was found for 10 items, the ranking of the three countries differing across these items.

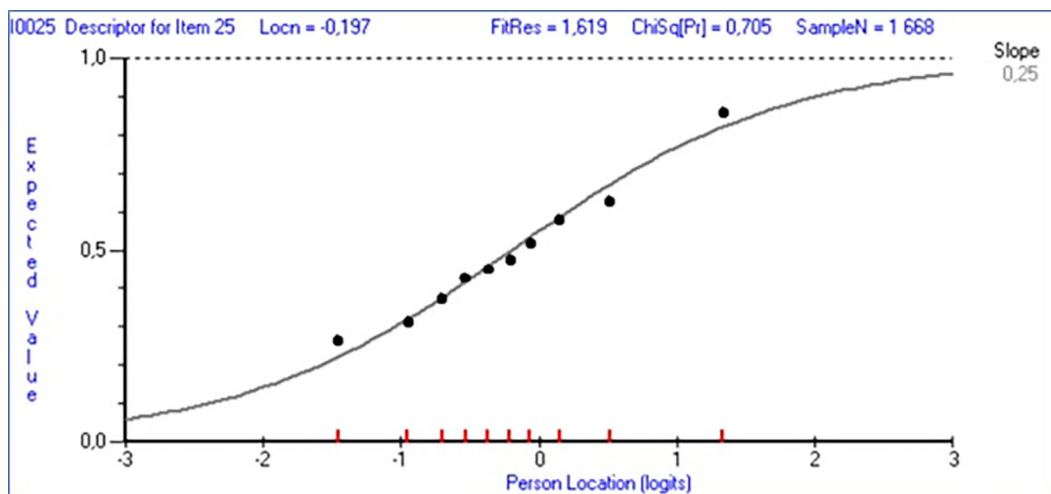
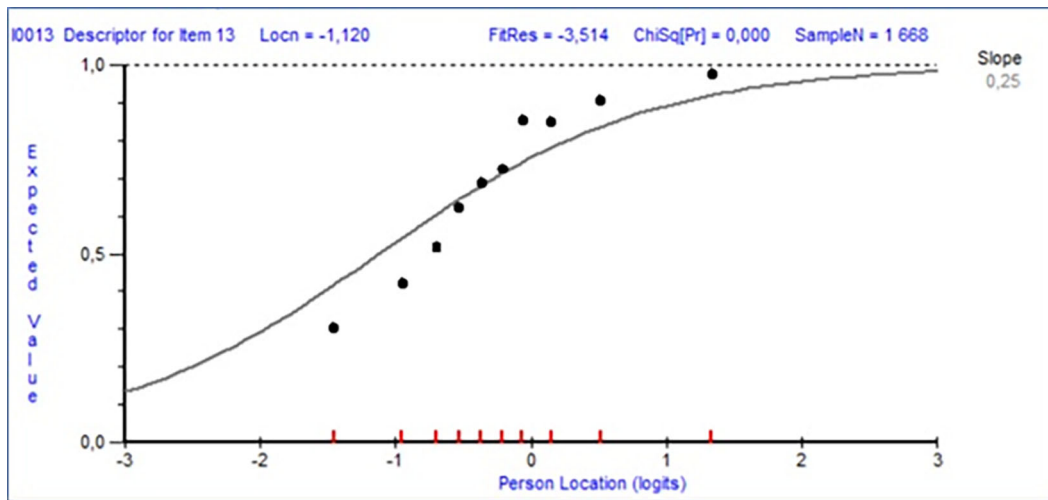
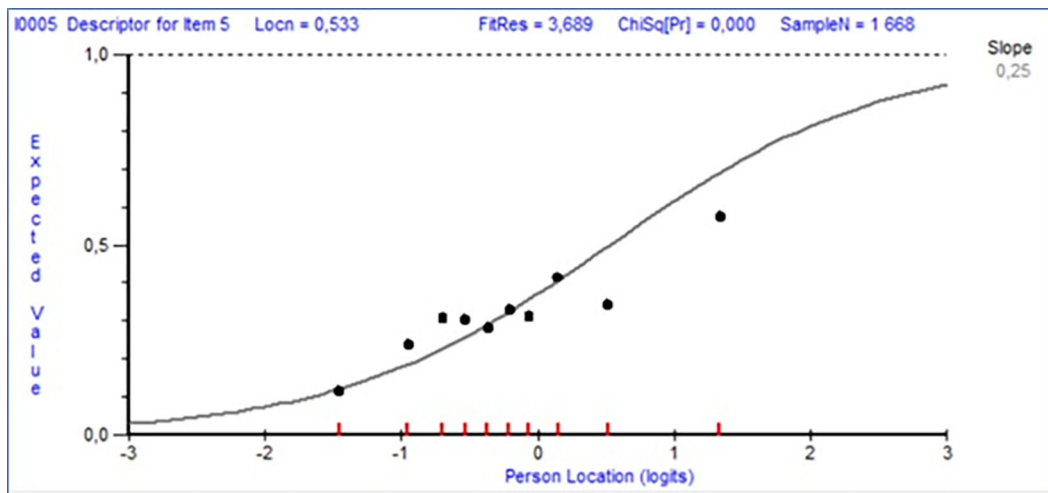


Figure 2. Example of ability-item with satisfactory fit.





**Figure 3.** Example of ability-item with classic over-discrimination.



**Figure 4.** Example of ability-item with classic under-discrimination.

There was no DIF by gender, age, or country in the analysis of the Likert item-set.

In the Likert item-set, two items were found to be slightly over-discriminating and were therefore considered acceptable. The remaining items showed very good fit.

When exploring the ordering of the thresholds, we found that the three Likert items evaluating intended behaviour were disorganized. A reanalysis of these suggest that these could be improved by dichotomising the response options. The four items evaluating self-efficacy showed a good fit.

### Test of unidimensionality

In the analysis of the ability item-set, 8% of the T-tests were significant.

The magnitude of multidimensionality in Likert-items were found satisfactory at 5% and considered to be unidimensional.

### Local dependency

There were no item-pairs correlations above 0.2 of the average value in any of the item-sets, suggesting no important redundancy.

### Revision of the questionnaire

The outcome measure to be used in the final trial was reduced to include only two MCQs for each Key Concept to be assessed. We removed the ability-items with suboptimal fit. Since the Likert-items were all found to have good fit, these remained unchanged.

The revised outcome measure has been published as extended data.<sup>37</sup>

### Discussion

Overall, both item-sets were found to have good fit to the Rasch model and suitable for our target audience. The reliability of both item-sets was also good. Observations of the individual item and person fit provided us with guidance on how to improve the design and administration of the two item-sets.

When observing each individual item's fit to the Rasch model in the ability item-set, we identified some items that could be removed to improve the questionnaire. Of 27 ability items, three had differential item functioning by age or gender of which only one of these were highly problematic (non-uniform). As expected, some items also showed differential item functioning by country. Possible explanations for this may be that there are differences in cultural beliefs or because there are differences in the curricula taught in schools. Considering that the differential item functioning by country was uniform and that we are planning to use the outcome measure in randomised trials comparing effects between comparison groups in each specific context, this was not considered to be a concern for our purposes. We also identified some items with poor measurement properties by observing the item characters curves. Taken together with the item showing non-uniform DIF, these were considered for removal from the final outcome measure to be used in our upcoming trial.

In the analysis of the Likert item-set, two issues were identified that we needed to address. Three items measuring intended behaviour showed disordered response categories, furthermore we identified a high number of people with extreme values. This can be an indication that some of the respondents had difficulty answering these questions. As noted in the methods, we observed that some people in the studied contexts were unfamiliar with intended behaviour and self-efficacy questions. The results from this study suggested that we need to plan carefully for how this item-set is administered and ensure that people are adequately instructed about the format and purpose of these questions. The results also suggested that we should either redesign the attitude items so that the response options are dichotomized (with three response options instead of five) or dichotomise the answers by collapsing the response options in the analysis following the trial. We did the latter in the trial of the IHC primary school resources by combining likely (or difficult) and very likely, and combining unlikely, very unlikely, and 'don't know'.<sup>26</sup>

We found no important redundancy in the item-sets (dependency between item pairs), and both item-sets appear to measure only one underlying trait (unidimensionality). The ability item-set had a somewhat higher percentage of T-tests above the statistical threshold of 5%.<sup>24</sup> Considering that this is the first time we have observed this in one of the many Rasch analyses we have done on instruments developed from the Claim Evaluation Tools item bank, we considered the magnitude of unidimensionality observed in the ability item-set acceptable.<sup>12-14</sup>

The overabundance of unreliable treatment claims that accompanied the COVID-19 pandemic has highlighted the need for facilitating critical thinking as an important public health initiative.<sup>5</sup> This is essential to protect people against unreliable treatment claims and enable them to make informed treatment choices.

Health literacy is defined in many ways, but typically includes the ability to think critically (sometimes referred to as critical health literacy).<sup>27,28</sup> A conceptual framework is helpful when developing assessment tools.<sup>29</sup> Health literacy is often measured using self-report.<sup>30</sup> Furthermore, many of the health literacy instruments available aim to capture other domains of health literacy such as functional and social literacy.<sup>30,31</sup> In addition to measuring perceptions of one's own abilities (self-report or self-efficacy), it is important to measure abilities objectively (performance). The association between self-report and performance is not straightforward.<sup>32</sup> The Health Literacy Tool shed, a database of health literacy measures has indexed 16 instruments evaluating an aspect of health literacy intended for adolescents using an objective measurement of performance, of which eight are available in English.<sup>30</sup> The Claim Evaluation Tools have a narrower scope than most of these and focusses on one critical skill, the ability to assess treatment claims and make informed treatment choices. Although these instruments can provide information about people's *general* health literacy skills, applying a more specific assessment tool in, for example, mapping studies, makes it easier to design interventions targeting the specific gaps identified.

### Strengths and limitations

One limitation of this study is that the adult population included more people with higher education than the general population in each of these three settings. Thus, the test might be more difficult for people with less education. However,

although participants with higher education are somewhat more likely to answering the ability questions correctly, there does not seem to be a strong association.<sup>14,33</sup> Another limitation is that the findings of this study are exclusive to the three Eastern African countries, and the validity and reliability in other contexts are uncertain. The item-sets validated in this study should therefore undergo further psychometric testing if used elsewhere.

The strategy of using pilot testing and a Rasch analysis have been found to be a robust method for developing measurement tools in several contexts.<sup>10–13</sup> An important strength of this study is that we used explicit and transparent methods, following the principal steps recommended for Rasch analysis.<sup>21–23</sup> Another strength is that we were able to recruit enough people despite the fact all three countries were burdened by the pandemic during the data collection. The results of this study and subsequent design of the questionnaire based on these results ensures that both the ability and Likert item-sets are a valid and reliable outcome measure for the randomised trials of the IHC lower secondary school intervention in all three countries.

## Conclusion

To our knowledge, this is the first measurement tool developed for measuring ability, intended behaviours, and self-efficacy for critical thinking about treatments in Kenya and Rwanda, as well as in Uganda. The two item-sets we evaluated in this study were found to be reliable and to have satisfactory measurement properties.

The findings from our analysis were used to redesign and improve the ability item-set. The results also informed guidance for how the Likert item-set should be administered and analysed.

## Data availability

### Underlying data

Zenodo: Critical thinking about treatment effects in Eastern Africa. Data set uncoded. [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.7680780>.<sup>34</sup>

The project contains the following underlying data:

- data-209546-2021-12-22-1147-utf\_final\_eclaim\_rasch\_2021.xlsx. (Raw data electronically collected - adults and students).
- data-237440-2021-12-22-1155-utf\_pilot Rwanda\_Rasch\_2021.xlsx. (Raw data collected from paper-based questionnaires used in the pilot survey - students).

### Extended data

Zenodo: Study protocol: Assessment of validity and reliability of a questionnaire based on the Claim Evaluation Tools Item bank in Uganda, Kenya and Rwanda. <https://doi.org/10.5281/zenodo.7680616>.<sup>35</sup>

The project contains the following extended data:

- Protocol\_Claim\_Choice 2021 23 03.docx.(2).pdf. (Study protocol)

Zenodo: Critical thinking about treatment effects in Eastern Africa. The Critical Thinking about Health test (before Rasch analysis). Zenodo. <https://doi.org/10.5281/zenodo.7756037>.<sup>36</sup>

The project contains the following extended data:

- Critical thinking about treatments test – Vis - Nettskjema.pdf. (Original test validated as part of this study).

Zenodo: Critical thinking about treatment effects in Eastern Africa. The Critical Thinking about Health test. <https://doi.org/10.5281/zenodo.7680606>.<sup>37</sup>

The project contains the following extended data:

- Test\_CHOICE\_final\_with literacy and userexperience\_march\_2022\_FORMATD.pdf. (Final revised test).

## Reporting guidelines

Zenodo: STROBE checklist for ‘Critical thinking about treatment effects in Eastern Africa: development and Rasch analysis of an assessment tool’. <https://doi.org/10.5281/zenodo.7680586>.<sup>38</sup>

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) (CC-BY 4.0).

## Acknowledgements

We would like to thank Sarah Rosenbaum for providing her expertise in designing the questionnaire. Furthermore, we would like to thank the rest of Informed Health Choices team for their valuable feedback and discussion in planning and conducting this study. We are also very grateful for all the secondary school students and adults who took time to contribute to this study and to the ministry of education and school administration for allowing students participation.

## References

- Mian A, Khan S: **Coronavirus: the spread of misinformation.** *BMC Med.* 2020; **18**(1): 89.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Oxman M, Larun L, Pérez Gaxiola G, *et al.*: **Quality of information in news media reports about the effects of health interventions: Systematic review and meta-analyses.** *F1000Res.* 2022; **10**(433): 433.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Brownlee S, Chalkidou K, Doust J, *et al.*: **Evidence for overuse of medical services around the world.** *Lancet.* 2017; **390**(10090): 156–168.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Glasziou P, Straus S, Brownlee S, *et al.*: **Evidence for underuse of effective medical services around the world.** *Lancet.* 2017; **390**(10090): 169–177.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- The Lancet Infectious D: **The COVID-19 infodemic.** *Lancet Infect. Dis.* 2020; **20**(8): 875.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- The Lancet R: **Going viral: misinformation in the time of COVID-19.** *Lancet Rheumatol.* 2021; **3**(6): e393.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Oxman A, Chalmers I, Austvoll-Dahlgren A, *et al.*: **Key Concepts for assessing claims about treatment effects and making well-informed treatment choices.** *F1000Res.* 2019; **7**(1784): 1784.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- 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, 1-year follow-up: a cluster-randomised trial.** *Trials.* 2020; **21**(1): 27.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Semakula D, Nsangi A, Oxman AD, *et al.*: **Effects of the Informed Health Choices podcast on the ability of parents of primary school children in Uganda to assess the trustworthiness of claims about treatment effects: one-year follow up of a randomised trial.** *Trials.* 2020; **21**(1): 187.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Austvoll-Dahlgren A, Guttersrud O, 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**(5): e013185.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Austvoll-Dahlgren A, Semakula D, Nsangi A, *et al.*: **Measuring ability to assess claims about treatment effects: the development of the ‘Claim Evaluation Tools’.** *BMJ Open.* 2017; **7**(5): e013184.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Perez-Gaxiola G, Austvoll-Dahlgren A: **Validación de un cuestionario para medir la habilidad de la población general para evaluar afirmaciones acerca de tratamientos médicos.** *Gac. Med. Mex.* 2018; **154**(4): 480–495.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Wang Q, Austvoll-Dahlgren A, Zhang J, *et al.*: **Evaluating people’s ability to assess treatment claims: Validating a test in Mandarin from Claim Evaluation Tools database.** *J. Evid. Based Med.* 2019; **12**(2): 140–146.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Dahlgren A, Furuseth-Olsen K, Rose C, *et al.*: **The Norwegian public’s ability to assess treatment claims: results of a cross-sectional study of critical health literacy [version 2; peer review: 1 approved, 1 approved with reservations].** *F1000Res.* 2021; **9**(179).  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Bloem EF, van Zuuren FJ, Koenen MA, *et al.*: **Clarifying quality of life assessment: do theoretical models capture the underlying cognitive processes?** *Qual. Life Res.* 2008; **17**(8): 1093–1102.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Choi SW, Cook KF, Dodd BG: **Parameter recovery for the partial credit model using MULTILOG.** *J. Outcome Meas.* 1997; **1**(2): 114–142.  
[PubMed Abstract](#)
- Linacre JM: **Sample size and item calibration stability.** *Rasch Measurement Transactions.* 1994; **328**.
- Clauser BE, Mazor KM: **Using Statistical Procedures to Identify Differentially Functioning Test Items.** *Educ. Meas. Issues Pract.* 1998; **17**(1): 31–44.
- Rogers HJ, Swaminathan H: **A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning.** *Appl. Psychol. Meas.* 1993; **17**(2): 105–116.  
[Publisher Full Text](#)
- Narayanan P, Swaminathan H: **Performance of the Mantel-Haenszel and Simultaneous Item Bias Procedures for Detecting Differential Item Functioning.** *Appl. Psychol. Meas.* 1994; **18**(4): 315–328.  
[Publisher Full Text](#)
- Tennant A, Conaghan PG: **The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?** *Arthritis Rheum.* 2007; **57**(8): 1358–1362.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Psychlab Group: **Introductory Rasch Analysis Using RUMM2030.** Psychometric Laboratory for Health Sciences: The Section of Rehabilitation Medicine University of Leeds; 2016.
- Rumm Laboratory Pty Ltd: **Displaying the RUMM2030 analysis.** Rasch unidimensional measurement model; 2015.
- Hagell P: **Testing Rating Scale Unidimensionality Using the Principal Component Analysis (PCA)/t-Test Protocol with the Rasch Model: The Primacy of Theory over Statistics.** *Open J. Stat.* 2014; **04**: 456–465.  
[Publisher Full Text](#)
- Andrich D: **Rasch Models for Measurement.** Beverly Hills: Sage Publications I; 1988.
- 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**(10092): 374–388.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Chinn D: **Critical health literacy: A review and critical analysis.** *Soc. Sci. Med.* 2011; **73**(1): 60–67.  
[Publisher Full Text](#)
- 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**(6): e020080.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
29. **COSMIN Taxonomy of Measurement Properties.**  
[Reference Source](#)
  30. **Health Literacy Tool Shed: A database of health literacy measures.**  
[Reference Source](#)
  31. Nguyen TH, Paasche-Orlow MK, McCormack LA: **The State of the Science of Health Literacy Measurement.** *Stud. Health Technol. Inform.* 2017; **240**: 17–33.  
[PubMed Abstract](#)
  32. Kiechle ES, Bailey SC, Hedlund LA, et al.: **Different Measures, Different Outcomes? A Systematic Review of Performance-Based versus Self-Reported Measures of Health Literacy and Numeracy.** *J. Gen. Intern. Med.* 2015; **30**(10): 1538–1546.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  33. Sørensen K, Pelikan JM, Röthlin F, et al.: **Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU).** *Eur. J. Pub. Health.* 2015; **25**(6): 1053–1058.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  34. Dahlgren A: **Critical thinking about treatment effects in Eastern Africa.** Data set uncoded. Data set. *Zenodo*. 2023.  
[Publisher Full Text](#)
  35. Dahlgren A, Semakula D, Oxman A, et al.: **Study protocol: Assessment of validity and reliability of a questionnaire based on the Claim Evaluation Tools Item bank in Uganda, Kenya and Rwanda.** Dataset. *Zenodo*. 2023.  
[Publisher Full Text](#)
  36. Dahlgren A: **Critical thinking about treatment effects in Eastern Africa. The Critical Thinking about Health test (before Rasch analysis).** *Zenodo*. 2023.  
[Publisher Full Text](#)
  37. Dahlgren A: **Critical thinking about treatment effects in Eastern Africa. The Critical Thinking about Health test.** *Zenodo*. 2023.  
[Publisher Full Text](#)
  38. Dahlgren A: **Critical thinking about treatment effects in Eastern Africa: development and Rasch analysis of an assessment tool. STROBE checklist.** *Zenodo*. 2023.  
[Publisher Full Text](#)

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

**F1000Research**