Informed Health Choices

Measuring ability to assess claims about treatment effects in English and Luganda: evaluation of multiple-choice questions from the "Claim Evaluation Tools" database using Rasch modelling

Semakula D A et al.

Working paper, 17. March 2017

Colophon

Title Measuring ability to assess claims about treatment effects in English and Luganda: evaluation of multiple-choice questions from the "Claim Evaluation Tools" database using Rasch modelling

Authors Daniel Semakula
Allen Nsangi
Andrew D. Oxman
Nelson K. Sewankambo
Øystein Guttersrud
Astrid Austvoll-Dahlgren

Corresponding author(s) Astrid Austvoll-Dahlgren
astrid.austvoll-dahlgren@fhi.no
Norwegian Institute of Public Health
PO Box 4404, Nydalen
N-0403 Oslo, Norway

Keywords evidence-based medicine, shared decision making, health literacy, outcome measurement, multiple-choice, patient education, Rasch analysis

Citation Semakula D, Nsangi A, , Oxman AD, Sewankambo NK, Øystein G, Austvoll-Dahlgren A. Measuring ability to assess claims about treatment effects in English and Luganda: evaluation of multiple-choice questions from the "Claim Evaluation Tools" database using Rasch modelling. Informed Health Choices Working Paper, 2017.

Date March 2017

Plain language summary

The Informed Health Choices (IHC) project has developed learning resources to teach primary school children and their parents to assess claims about the effects of treatments (any action intended to improve health). As part of this project, we have developed a database of multiple-choice questions to measure an individual's ability to assess treatment claims. Each question is designed to measure an individual's understanding and ability to apply one of 34 Key Concepts that are important for people to understand and apply when assessing claims about treatment effects.

In previous studies, we showed that tests (sets of multiple-choice questions from the database) provided a valid and reliable measure of an individual's ability to assess treatment claims. However, the tests were difficult for our target groups, possibly because of low literacy and poor English skills. The purpose of this study was to evaluate two sets of multiple-choice questions (tests), selected for use in randomised trials of IHC learning resources in Uganda, administrated as oral tests in Luganda and as written tests in English.

We translated the previously validated and revised questions from English to Luganda. Each test included 26 multiple-choice questions evaluating individuals' abilities to apply 13 Key Concepts (two questions per concept). These two tests were then administrated as oral tests in Luganda and as written tests in English. The two tests were administered to 1617 people in Uganda, including children (<10) and adults, with and without relevant training. Each of the participants completed one of the two tests either in English as a written test or Luganda as an oral test.

We found that overall the tests were valid and reliable, and we could rule out any important differences between administering the tests in different ways and languages. The orally administered Luganda questions were better suited for our target groups, suggesting this helped address difficulty due to low literacy and poor English skills.

Based on the findings from this study, we chose the multiple-choice questions with the best fit to the model used to evaluate the questions and developed the final test to be used as outcome measures in trials of the IHC learning resources in Uganda. These 26 questions can be administrated in either of two languages and modes of administration to assess an individual's ability to understand and apply the 13 Key Concepts.

Abstract

Background: The Informed Health Choices (IHC) project has developed learning resources to teach children and their parents to assess claims about the effects of treatments (any action intended to improve health). As part of this project, we have developed the Claim Evaluation Tools database, which contains multiple-choice questions (MCQs) to measure an individual's ability to assess treatment claims. Each MCQ is designed to measure an individual's understanding and ability to apply one of 34 Key Concepts that are important for people to understand and apply when assessing claims about treatment effects. In a previous study, we used Rasch analysis to evaluate MCQs from the database to be used in randomised trials of the IHC learning resources in Uganda. That study included 88 MCQs addressing 22 Key Concepts. We found that overall the MCQs fit the Rasch model and the tests were reliable. However, the MCQs were difficult for the target groups, possibly due to low literacy and poor English skills.

Objectives: To evaluate two sets of MCQs, selected for use in randomised trials of IHC learning resources in Uganda, administrated as oral tests in Luganda and as written tests in English using Rasch analysis.

Methods: We translated the previously validated and revised MCQs from English to Luganda and created two sets of MCQs or "tests". Each test included 26 MCQs evaluating individuals' abilities to apply 13 Key Concepts (two MCQs per concept). These two tests were then administrated as oral tests in Luganda and as written tests in English. We scored all responses dichotomously, as correct or incorrect. We explored summary and individual fit statistics using the RUMM2030 analysis package. Potential differential item functioning (DIF) was explored for age, gender, and language/mode of administration. We used SPSS to perform distractor analysis.

Results: We found that overall the MCQs administered in English fit the Rasch model and the tests were reliable, supporting our previous findings. The MCQs administrated in Luganda had more limitations, and five required revision. None of the MCQs showed DIF by gender, and only two showed DIF by language. The orally administered Luganda MCQs had better targeting to the study participants, suggesting this helped address difficulty due to low literacy and poor English skills.

Conclusion: We could rule out any important differences between administering the tests in different ways and languages. Based on the findings from this study, we chose the MCQs with the best fit to the Rasch model and developed the final test to be used as outcome measures in trials of the IHC learning resources in Uganda. These 26 MCQs can be administrated in either of two languages and modes of administration to assess an individual's ability to understand and apply 13 Key Concepts.

Background

The Informed Healthcare Choices (IHC) project aims to help people assess treatment claims and make informed health choices. The project has developed primary school resources and a podcast series to improve the ability of children and their parents to assess claims about treatment effects. We have piloted these resources in Uganda, Kenya, Rwanda, and Norway. We will test the effects of the resources in randomized trials in Uganda [1,2].

The first step in the IHC project was to identify the Key Concepts people need to know to be able to assess treatment effects [3]. This resulted in an initial list of 32 Key Concepts that serves as a syllabus for designing learning resources. Two additional concepts were subsequently added to the Key Concepts list. This was also the starting point for the IHC learning resources. We present a short list of the Key Concepts in Table 1. The IHC primary school resources teach 12 of the Key Concepts to primary school children. These resources include a textbook and a teachers' guide. The textbook includes a comic, exercises and classroom activities. The IHC podcast for the parents of primary school children covers nine of the Key Concepts. Each episode of the podcast includes a short story with an example of a treatment claim, a simple explanation of a concept used to assess that claim, another example of a claim illustrating the same concept, and its corresponding explanation. Eight Key Concepts are covered by both the IHC primary school resources and the podcast, so that together they address a total of 13 Key Concepts (Table 1).

Table 1 *Key Concepts*

Included concepts		Key Concepts
School	Podcast	
		1. Claims: are they justified?
1	1	1.1 Treatments may be harmful
2	2	1.2 Personal experiences or anecdotes (stories) are an unreliable basis for assessing the effects of most treatments
	3	1.3 An 'outcome' may be associated with a treatment, but not caused by the treatment
3	4	1.4 Widely used treatments or treatments that have been used for a long time are not necessarily beneficial or safe
4		1.5 New, brand-named, or more expensive treatments may not be better than available alternatives
5	5	1.6 Opinions of experts or authorities do not alone provide a reliable basis for deciding on the benefits and harms of treatments
6		1.7 Conflicting interests may result in misleading claims about the effects of treatments
		1.8 Increasing the amount of a treatment does not necessarily increase the benefits of a treatment and may cause harm
		1.9 Earlier detection of disease is not necessarily better
		1.10 Hope or fear can lead to unrealistic expectations about the effects of treatments
		1.11 Beliefs about how treatments work are not reliable predictors of the actual effects of treatments

Included concepts		Key Concepts
School	Podcast	
		1.12 Large, dramatic effects of treatments are rare
		2. Comparisons: are they fair and reliable?
7	6	2.1 Evaluating the effects of treatments requires appropriate comparisons
8	7	2.2 Apart from the treatments being compared, the comparison groups need to be similar (i.e. 'like needs to be compared with like')
		2.3 People's outcomes should be counted in the group to which they were allocated
		2.4 People in the groups being compared need to be cared for similarly (apart from the treatments being compared)
9		2.5 If possible, people should *not* know which of the treatments being compared they are receiving
		2.6 Outcomes should be measured in the same way (fairly) in the treatment groups being compared
		2.7 It is important to measure outcomes in *everyone* who was included in the treatment comparison groups
10	8	2.8 The results of single comparisons of treatments can be misleading
		2.9 Reviews of treatment comparisons that do not use systematic methods can be misleading
		2.10 Unpublished results of fair comparisons may result in biased estimates of treatment effects
		2.11 Results for a selected group of people within a systematic review of fair comparisons of treatments can be misleading
		2.12 Relative effects of treatments alone can be misleading
		2.13 Average differences between treatments can be misleading
11		2.14 Small studies in which few outcome events occur are usually not informative and the results may be misleading
		2.15 The use of p-values to indicate the probability of something having occurred by chance may be misleading; confidence intervals are more informative
		2.16 Saying that a difference is statistically significant or that it is not statistically significant can be misleading
		2.17 A lack of evidence is not the same as evidence of "no difference"
		3. Choices: make informed choices
		3.1 A systematic review of fair comparisons of treatments should measure outcomes that are important
		3.2 A systematic review of fair comparisons of treatments in animals or highly selected groups of people may not be relevant
		3.3 The treatments evaluated in fair comparisons may not be relevant or applicable
		3.4 Well done systematic reviews often reveal a lack of relevant evidence, but they provide the best basis for making judgements about the certainty of the evidence
12	9	3.5 Decisions about treatments should not be based on considering only their benefits

The Claim Evaluation Tools database was developed to meet the needs of people interested in evaluating the ability of individuals to assess treatment claims and make informed health choices [4,5]. It includes multiple-choice questions (MCQs) to assess people's ability to apply the Key Concepts and assess claims about treatment effects.

From this database, researchers, teachers and others can select those MCQs that are relevant for specific populations and purposes. The MCQs include scenarios intended to be relevant across different contexts. They can be used for children (from ages 10 and up) and adults, including both patients and health professionals. In another paper, we have described the iterative development of the Claim Evaluation Tools database, including qualitative and quantitative feedback from experts and end-users in Uganda, Kenya, Rwanda, Norway, the United Kingdom, and Australia [5]. Each MCQ addresses one Key Concept.

In a previous study, validating four subsets of MCQs from this database in English as written tests, the MCQs were found to have satisfactory construct validity and reliability [6]. We also concluded that the MCQs seemed to function in the same way across subgroups of participants [6]. However, this first study also suggested that the MCQs were difficult for some people in our target groups. Although the ability to assess claims about treatment effects is generally low in many populations, we identified two additional barriers in the Ugandan setting that warranted attention; low literacy and poor English skills.

Therefore, we wanted to determine if it was possible to administer the MCQs as an oral test in Luganda. Even though English is the official language in Uganda, Luganda is the first language to many people in Central Uganda. To rule out any differential item functioning (item bias) caused by the two different languages and modes of administration, we compared the results from the Luganda MCQs administered orally with written English MCQs. For these tests we selected MCQs with best fit to the Rasch model based on findings from the first Rasch analysis [6]. Some of these items were also revised to simplify the text in the scenarios or by removing response options with poor fit to the Rasch model.

The findings of this study will inform the development of the primary outcome measure to be used in the randomised trials evaluating the IHC learning resources for primary school children and their parents in Uganda.

Methods

Objective

The objective of this study was to evaluate two sets of MCQs (tests), selected for use in randomised trials of IHC learning resources in Uganda, administered as oral tests in Luganda and as written tests in English using Rasch analysis.

Selection of MCQs for the two tests

For this study, we only tested MCQs that addressed the 13 Key Concepts that were targeted in the IHC learning resources and which had good fit to the Rasch

model based on our previous study [6] (Table 1). Since several MCQs are available for each Key Concept, we wanted to include most of these for this second validation. This was important because if an MCQ was judged to have poor fit to the Rasch model based on this second validation - for example, if we found important differences in an MCQ between the English and Luganda translations - we would have more than one candidate for each MCQ to choose from when creating the final test for our trials.

Translation

We translated the MCQs in 3 sequential steps:

1) The investigators, who are Luganda speakers and health researchers, translated the English MCQs to Luganda with the help of a Luganda language teacher. They read all the instructions, scenarios, questions and response options, and then translated them one at a time being careful to retain meaning.

2) A second Luganda language teacher reviewed the translations using the English tests as the reference document.

3) A third teacher was given only the Luganda translation and asked to back-translate the tests to English. To resolve inconsistencies between the back-translated version and the original English version, we asked at least five members of the public what meaning they derived from the question in Luganda. Adjustments were then made to the questions based on a consensus among the investigators and the translation support team, informed by the responses from the public.

Preparation of the audio version of the tests

Following the translation of the English tests to Luganda, we audio-recorded the Luganda tests verbatim. A radio presenter who was eloquent in Luganda read aloud all the text (instructions, questions and response options). We user-tested the first version of the audio recorded tests with a convenience sample of 15 Luganda speaking members of public to find out how suitable the audio tests were to the target group. This was done using a concurrent think aloud technique and a semi-structured interview guide. We adjusted the audio tests based on the findings of the user testing until we were satisfied that the audio tests were understandable and easy to use. Based on the findings of the user-testing, we decided to repeat every question twice to allow users sufficient opportunity to reflect on the question and understand it before responding. We also included sufficient pauses between questions to allow the user time to respond before

the next question played. The final audio tests were one hour and 15 minutes. We included faint instrumental background music to help reduce boredom.

Participants

We used purposeful sampling, including both children and adults, to explore item bias (differential item functioning) associated with age. We also made sure that we had an equal gender distribution and included a mix of people with and without relevant training. There is no consensus on the sample size needed to perform a Rasch analysis [7]. This is a pragmatic judgement that takes account of the number of items evaluated and the statistical power needed to identify item bias resulting from relevant background factors. Since we intended to test many items, we did not consider it feasible to include these in a single test, and split the items into two sets or "tests". This resulted in a total of four sets of questions, two in Luganda and two in English. We judged that at least 250 respondents per set were needed for this study.

Administration of the tests

We installed the audio tests on portable media players with speakers. Research assistants together with the participants found a quiet place where they sat and the assistants played the tests to the participants individually, recording each of their responses on the corresponding question on the Luganda paper test. The research assistants had no prior interaction with the content of tests and except for the training they received on how to administer them, they did not know the correct answers to the questions. Participants could pause and resume the test at will, like they would do with the paper test. At the end of the interaction the research assistant thanked the participant for taking part in the process.

Rasch analysis and item response theory

Rasch analysis relies on item response theory, which is a paradigm for designing and testing measurement instruments used for assessing people's abilities, attitudes, and other attributes. Rasch analysis is used to check the degree to which scoring and summing-up across items is defensible in the data collected [8,9]. It is a unified approach to address important measurement issues required for validating an outcome measure, such as a scale or a test, including testing for internal construct validity (by testing for multidimensionality), invariance of the items (item-person interaction), and item bias (differential item function) [9].

When developing outcome measurements, Rasch analysis provides an excellent basis for revising individual items informed by judgements about misfit to the

Rasch model [10]. By identifying misfit to the model, individual items can be repaired by removing sub-optimal response options, by collapsing response options, or by removing suboptimal items altogether. In this way, the Rasch analysis represents a dynamic approach to achieving construct validity.

Preparation of the dataset

We used EpiData data entry software version 3.1 [11] to create the electronic database for the two sets of tests, into which we entered and stored the data. To ensure accuracy during data entry, two different individuals entered each test twice. We resolved discrepancies in data entry by checking the original paper test and, if there were differences in interpretation, by consensus together between two of the investigators (DS and AN). After data entry was completed, we exported the data to Excel and cleaned them before entering them into RUMM2030 for Rasch analysis [12]. For this analysis, we scored all responses to MCQs dichotomously, as correct or incorrect.

Item-person interaction and reliability

Initial analysis of the MCQs was performed by using the summary statistic function in RUMM2030 to explore the item-person interaction [12].

In RUMM2030, the *mean item location* of the items is always centralized as "0" on a logit scale forming a normal distribution. The *persons' location* provides information about the *targeting* of the test. A higher value than "0" would indicate that the ability of the respondent was higher than the test (an easy test), and a lower score would indicate that the mean person ability was lower than the test (a hard test) [13].

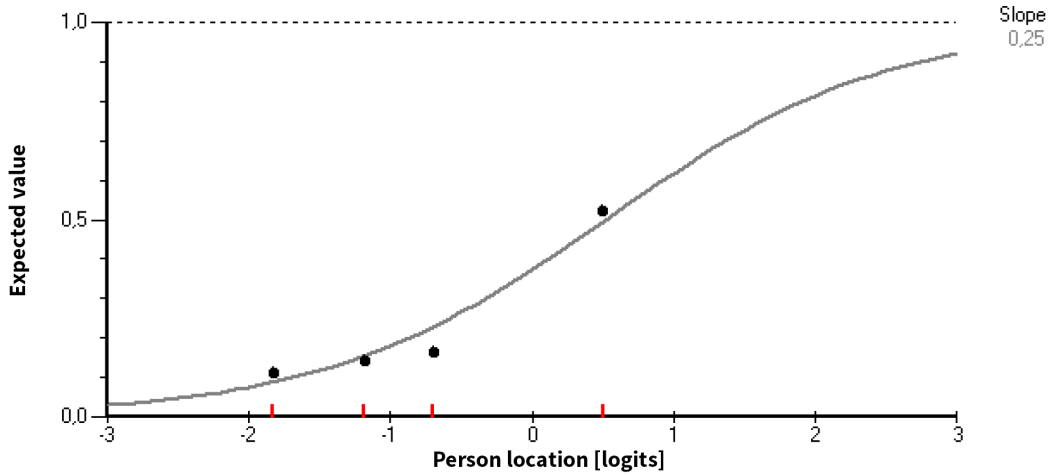
The overall *Item* and *Person Fit Residual* statistics assess the degree of divergence (or residual) between the expected and observed data for each person item when summed for all items and all persons respectively. In RUMM2030 this is reported as an approximate z-score, representing a standardized normal distribution [13]. Ideally, item fit and person fit should have a mean of zero and a SD of one [9,13].

We calculated Cronbach's Alpha as a measure of the reliability of each set of items. We considered a value of 0.7 or higher to be adequate for this. Cronbach's Alpha is only available if there are no missing data. We counted missing responses as "incorrect" [13].

Individual person and item fit

For each item, we visually inspected fit to the graphical representation of the specific Rasch model applied – the Item Characteristic Curve. An example is shown in Figure 1. Using the chi-square goodness of fit test, we formally compared the observed number of correct responses to the theoretically expected number given by the Rasch model. We set the significance level at 5%. The chi-square test statistic obtained for each item was compared to the critical value for the given degrees of freedom. When the test statistic is larger than the critical value the *p*-value will be below 5% and null hypothesis (that the item fits the Rasch model expectations) should be rejected. Using Bonferroni adjustment, the p-values were adjusted ($p = 0.05/k$) for the number of significance tests (k) carried out (one for each item).

Figure 1 *The Item Characteristic Curve*



We also performed distractor analysis using SPSS. The latter is particularly useful in developing and revising multiple-choice questions, because it may identify response options (distractors) that are not working as intended and can subsequently be deleted or revised.

Testing for multidimensionality and response dependency

Local dependency is a requirement of the Rasch model, and assumes that there is one construct explaining how the items are related to each other and that the items are conditionally independent [9,14]. This can be explored by testing for response dependency and by exploring potential multi-dimensionality of the data [15,16]. We explored possible dimension violations of local independence applying the PCA/t-test procedure computing paired t-tests using two sub-sets of items from each item set. The hypothesis of a unidimensional scale is weakened when the proportion of individuals with statistically significant differences in ability estimates on a pair of subscales exceeds 5% [17]. We also inspected

the residual correlation matrix estimated in RUMM2030 [18]. We considered residual correlations above 0.3 as indicators of response dependence between items [19].

Results

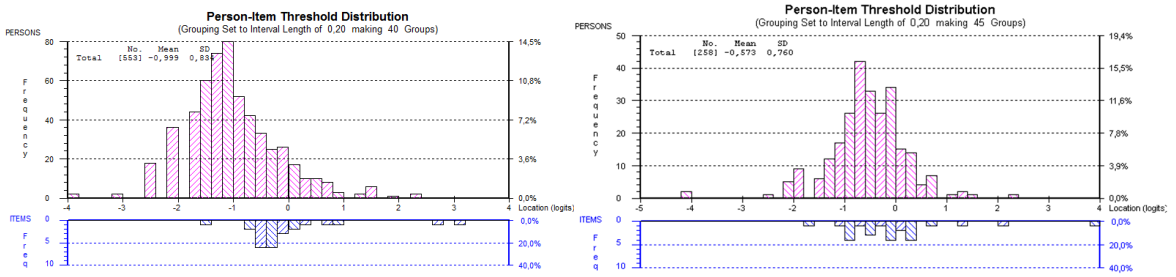
Overall, we recruited 1617 people for this validation. The two English samples each consisted of 553 people (of which 585 were children), and the Luganda samples consisted of 258 (Test 1) and 253 (Test 2) people respectively.

Item-Person Interaction and reliability

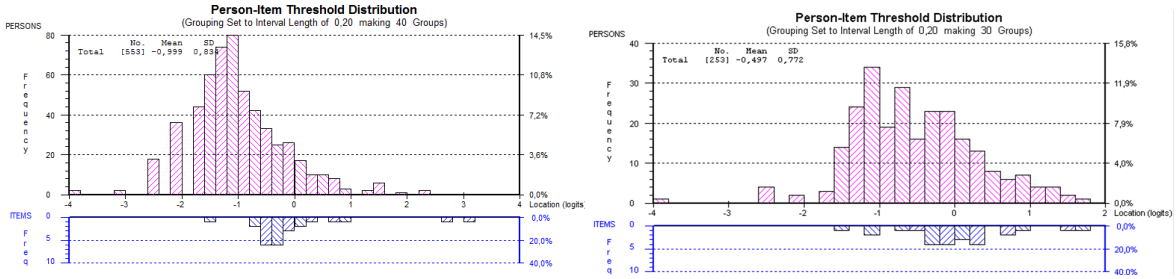
The Person-Item distribution for each test and by language/ mode of administration is presented in Figure 2. The bars on the upper part of these graphs represent ability groups of respondents in a normal distribution, the bars under the line represents the item locations (difficulty). When items have the same difficulty, they are located in the same place (building on each other). Consequently, it can be observed from these graphs that the items are well distributed but with some clustering around zero. Furthermore, there are a few extremely difficult items and some very easy items.

Figure 2 *Person-item distribution English version (left) and Luganda version (right)*

Test 1



Test 2



The ability groups are also clustered around zero, but with most located under zero and very few with a person location over zero. Indicating that overall this is a somewhat difficult test, and that based on this test, and particularly for the English tests, the MCQs may not be sensitive enough to discriminate well be-

tween the lowest ability groups. Although the patterns are similar, the Luganda tests show better targeting and the respondents were more likely to answer correctly.

For the English written test, the observed logits for sets 1 and 2 were -1.0 and -0.85 logits respectively. The reliability was satisfactory for both tests, with a Cronbach's Alpha of 0.7 for test 1 and 0.76 for test 2.

For the Luganda-oral tests, the observed logits for sets 1 and 2 were -0.57 and -0.49 logits. The reliability for both tests was slightly below what we considered satisfactory, with a Cronbach's Alpha of 0.6 for test 1 and 0.65 for test 2.

Individual item fit and differential item functioning

Individual item fit supported previous findings from the first Rasch analysis of the MCQ's in English, in that few MCQs under discriminated according to the Rasch model expectations. That is, there was not "statistically significantly" more variation in the item data than predicted by the Rasch model. In the Luganda tests, four MCQs in test 1 and three in test 2 had poor fit. We considered these items as candidates for revision or rejection.

Two MCQs in test 1 and four MCQs in test 2 displayed uniform DIF associated with age. Furthermore, three MCQs in test 1 and two MCQs in test 2 displayed DIF associated with type of administration/language (written English or oral Luganda). However, this DIF was uniform, which we considered to be acceptable. We did not observe DIF associated with gender.

Response dependency and unidemensionality

We did not observe any subset of dependent items bringing specific variance into the scale – systematic variance not modelled by the Rasch factor "ability to assess claims about treatment effects". Relying on principal component analysis of residuals and paired t-tests we found that the Rasch factor "ability" – one latent variable – could sufficiently "explain" the observed covariance between the MCQs.

Creation of the final test for the IHC trials in Uganda

For both the English and Luganda MCQs, we made a note of the MCQs with good fit and with lower difficulty based on the group mean z-score for each MCQ. The MCQs selected for the initial English tests all had good fit. More issues were identified in the Luganda MCQs – and we considered revising five of these. For three of these, we revised response options that seemed to attract those with high ability. This was not a problem for the English MCQs. We also reviewed two

MCQs with signs of uniform DIF based on language (English or Luganda) to check if the translated MCQs were too easy (1 MCQ) or too difficult (1 MCQ).

Based on these considerations, we created a test with 26 MCQs (2 for each of 13 Key Concepts) to be used as the primary outcome measure in randomised trials of the IHC primary school resources and the IHC podcast in Uganda.

Discussion

Overall, we found that the MCQs administered as a written test in English were reliable and with overall good fit to the Rasch model, supporting previous findings [6]. The MCQs administrated in Luganda had more limitations, but because of the large number of MCQs tested, we could choose MCQs that showed satisfactory fit after revising five of these. We identified only three MCQs with DIF based on language and mode of administration.

This was the first study evaluating MCQs from the Claim Evaluation Tools database that explored potential DIF by gender. We identified no such item bias. There was some evidence of DIF by age for six MCQs. However, this DIF was uniform and we considered this to be acceptable.

In our previous validation of the MCQs in English, the MCQs were found to be difficult [5]. Results of the present study suggest that the English tests were similar in difficulty to what was found in the previous analysis, where the observed logits for the four sets tested ranged from -0.81 to -1.15 [6]. However, the results from the oral Luganda tests showed better targeting, suggesting this helped address difficulty due to low literacy and poor English skills.

Based on these findings, we chose the MCQs with the best fit to the Rasch model and developed the final test to be used as the primary outcome measure in the trials evaluating the IHC primary school resources and podcast. This test can be administrated in two languages. In the primary school trial, the oral Luganda version will be administered to a sample of children in each participating school. In this way, we will test the extent to which literacy and having English as a second language might have affected the results using the written English test, which will be administered to all the children participating in the trial. In the podcast trial, participants will be able to choose whether they want to complete the written English version or the oral Luganda version.

Researchers in Norway, Mexico (Spanish), Germany, China and the UK are currently testing the MCQs in their settings. All MCQs in the Claim Evaluation Tools database are freely available for non-commercial use on request through the Testing Treatments interactive [20].

Conclusion

The findings from this second Rasch analysis, conducted in Uganda using two subsets of MCQs from the Claim Evaluation Tools database, confirms previous findings that the MCQs administered in English as a written test are reliable and with overall good fit to the Rasch model. The MCQs administrated orally in Luganda had more limitations, and some required revision. None of the MCQs showed bias by gender, and only two MCQs showed DIF based on language or mode of administration. The Luganda MCQs had better targeting for the study sample. Therefore, it seems likely that we have succeeded in addressing some of the challenges relating to literacy. Based on these findings, we created tests including MCQs with satisfactory fit to the Rasch model. These will be used as the primary outcome measure in randomised trials evaluating the effects of the IHC primary school resources and the IHC podcast in Uganda. The test will be administrated as a written test in English and as an oral test in Luganda.

Authors' Contributions

AA, AN, AO, DS, and NS conceptualised and planned this study. AN and DS collected the data with support from AA and NS. AN and DS entered the data. AA and ØG prepared the data files for the analysis, and ØG conducted the Rasch analysis. AA prepared the first draft of this manuscript and all the authors contributed to the final version.

Ethical approval

The research was approved by the Makerere University School of Medicine Research and Ethics Committee and the Uganda National Council for Science and Technology.

Acknowledgements

We thank Sarah Rosenbaum for help designing the test, and Kjetil Olsen for help with data entry. We are grateful to all the enthusiastic children, parents, teachers and journalists that contributed to this project.

Funding and competing interests

The IHC project is funded in part by the Research Council of Norway- GLOBVAC project 220603. The authors declare no conflicts of interests.

References

1. Nsangi A., Semakula D., Oxman M., Austvoll-Dahlgren A., Rosenbaum S., Kaseje M., et al. Evaluation of resources to teach children in low income countries to assess claims about treatment effects. Protocol for a randomized trial. Trials. In press.
2. Semakula D., Nsangi A., Oxman M., Austvoll-Dahlgren A., Rosenbaum S., Kaseje M., et al. Can an educational podcast improve the ability of parents of primary school children to assess the reliability of claims made about the benefits and harms of treatments: study protocol for a randomised controlled trial. Trials 2017; 18:31.3.
3. 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.
4. Austvoll-Dahlgren A, Nsangi A, Semakula D. Interventions and assessment tools addressing key concepts people need to know to appraise claims about treatment effects: a systematic mapping review. Syst Rev 2016; 5:215.
5. Austvoll-Dahlgren A, Semakula D, Nsangi A, Oxman A, Chalmers I, Rosenbaum S, et al. Measuring ability to assess claims about treatment effects: The development of the “Claim Evaluation Tools”. BMJ Open. In press.
6. Austvoll-Dahlgren A, Guttersrud G, Nsangi A, Semakula D, Oxman A; IHC Group. Measuring ability to assess claims about treatment effects: A latent trait analysis of the “Claim Evaluation Tools” using Rasch modelling. BMJ Open. In press.
7. Linacre J. Sample size and item calibration stability. Rasch Meas Trans 1994; 7:328.
8. Leonard M. Rasch promises: a layman's guide to the Rasch method of item analysis. Educ Res 1980; 22:188-92.
9. 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; 57:1358-62.
10. Rasch analysis. Accessed 20.03.2017. Available at: http://www.rasch-analysis.com/
11. Lauritsen JM, Bruus M. EpiData (version 3.1). A comprehensive tool for validated entry and documentation of data. Odense, Denmark: The EpiData Association, 2003-2005. www.epidata.dk

12. Andrich D, Lyne A, Sheridan B, Luo G. RUMM2030: Rasch Unidimensional Measurement Model software [computer program]. Perth: RUMM Laboratory, 2009.

13. Psylab Group. Introductory Rasch Analysis Using RUMM2030. Leeds: Section of Rehabilitation Medicine, University of Leeds, 2016.

14. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value Health 2004; 7(Suppl 1):S22-6.

15. Rasch G. Probabilistic models for some intelligence and achievement tests. Copenhagen: Danish Institue for Educational Research. Expanded Edition 1983. Chicago: MESA Press, 1960.

16. Marais I, Andrich D. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. J Appl Meas 2008; 9:200-15.

17. RUMM. Extending the RUMM2030 Analysis. 7th ed. Perth: RUMM Laboratory, 2009.

18. 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; 4:456-65.

19. Andrich D, Humphry SM, Marais I. Quantifying local, response dependence between two polytomous items using the Rasch model. Appl Psychol Meas 2012; 36:309–24.

20. Austvoll-Dahlgren A, Oxman AD, Chalmers I; Claim Evaluation Tools Database Working Group. Manual for preparing a test or questionnaire based on the Claim Evaluation Tools database. Version: 22.11.2016. Oslo: Informed Health Choices project, 2016. Available from: http://www.informedhealthchoices.org/wp-content/uploads/2016/08/Manual-tailoring-your-own-questionnaire_29112016.docx