BMJ Open

Measuring ability to assess claims about treatment effects: A latent trait analysis of the "Claim Evaluation Tools" using Rasch modelling

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2016-013185
Article Type:	Research
Date Submitted by the Author:	06-Jul-2016
Complete List of Authors:	Austvoll-Dahlgren, Astrid; Norwegian Institue Of Public Health, Guttersrud, Øystein; Norwegian Centre for Science Education, University of Oslo Nsangi, Allen; Makerere University College of Health Sciences, School of Medicine Semakula, Daniel; Makerere University College of Health Sciences, Oxman, Andrew; Norwegian Health Services Research Centre
<b>Primary Subject Heading</b>:	Patient-centred medicine
Secondary Subject Heading:	Research methods, Public health, Health policy
Keywords:	evidence based medicine, health literacy, outcome measurement, multiple-choice, Rasch analysis, patient education

Measuring ability to assess claims about treatment effects: A latent trait analysis of the “Claim Evaluation Tools” using Rasch modelling

Astrid Austvoll-Dahlgren, Øystein Guttersrud, Allen Nsangi, Daniel Semakula, Andrew D. Oxman, The IHC group*

Iain Chalmers

Leila Cusack

Claire Glenton

Tammy Hoffmann

Margaret Kaseje

Simon Lewin

Leah Atieno Marende

Michael Mugisha

Laetitia Nyirazinyoye

Kjetil Olsen

Matthew Oxman

Sarah Rosenbaum

Nelson K. Sewankambo

Anne Marie Uwitonze

Astrid Austvoll-Dahlgren (corresponding author)
astrid.austvoll-dahlgren@fhi.no
+47 41294057
Norwegian Institute of Public Health
BOKS 7004 St.Olavsplass
0130 Oslo, Norway

Øystein Guttersrud
oystein.guttersrud@naturfagsenteret.no
Norwegian Centre for Science Education, University of Oslo
Postboks 1106, Blindern 0317 Oslo, Norway

Allen Nsangi
nsallen2000@yahoo.com
Makerere University College of Health Sciences.
New Mulago Hospital Complex, Administration Building, Second Floor.
P.O.Box 7072, Kampala Uganda

Daniel Semakula
semakuladaniel@gmail.com
Makerere University College of Health Sciences.
New Mulago Hospital Complex, Administration Building, Second Floor.
P.O.Box 7072, Kampala Uganda

Andrew D. Oxman
oxman@online.no
Norwegian Institute of Public Health
BOKS 7004 St.Olavsplass
0130 Oslo, Norway

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

Abstract

Objectives: To describe the psychometric testing including Rasch analysis of the Claim Evaluation Tools in English speaking populations in Uganda and Norway.

Setting: We developed the Claim Evaluation Tools to evaluate members of the public's ability to assess treatment claims to be used in two randomised trials. The Claim Evaluation Tools consists of a battery of objective and flexible multiple-choice items.

Participants: We administrated four subsets of multiple-choice items to 1114 people, of which 685 were children and 429 were adults (including 171 health professionals). We scored all items dichotomously. Individual item fit statistics were estimated using pairwise maximum likelihood estimation available in the RUMM2030 analysis package and marginal maximum likelihood estimation available in the ConQuest4 analysis package. Possible dimension violations and response violations of local independence were studied using the PCA/t-test procedure and by inspecting the residual correlation matrix. We analysed differential item functioning for available person factors using the RUMM2030.

Results: The four subsets had satisfactory reliability, and most items did not over or under discriminate or function differently across subgroups of participants. There was no important response dependence, and one unidimensional latent trait was identified. The items had high difficulty. The spread (standard deviation) in the variable «ability» was as expected.

Conclusion: These results suggest that the items are reliable in the settings in which they have been tested. Items with sub-optimal fit will be considered deleted or repaired. We encourage further testing of the Claim Evaluation Tools in other contexts. All items will be available on request for non-commercial use through the website Testing Treatments interactive (www.testingtreatments.org).

Strengths and limitations of this study

- To our knowledge, this is the first psychometric testing including Rasch analysis of a set of evaluation tools that objectively measure people's ability to apply Key Concepts people need to know to assess claims about treatment effects

- We have used robust methods in two settings to test the reliability of the items, allowing for evidence informed revisions of any items with sub-optimal fit

- Our analysis suggest that the items are reliable in these settings and can be used for evaluating the effects of interventions to promote understanding of the Key Concepts

- This study provides evidence of the reliability of the items in two contexts, and the fit with the Rasch model when used in other settings is unknown.

- The items tested in this study was tested in English in English speaking populations, the reliability of the items in other languages remains to be tested.

Background

People are confronted with claims about treatment effects daily. This includes claims about the effects of changes in health behaviour, screening, other preventive interventions, therapeutic interventions, rehabilitation, and public health and health system interventions that are targeted at groups of people (1-4). Many of these claims are not based on evidence from fair comparisons of treatments, and many patients and health professionals alike don't have the necessary tools to assess the reliability of these claims (5-11). Being able to think critically and make informed decisions is essential for engaging patients in clinical decisions and citizens in policy decisions, and affects people's health and use of resources (10, 12-15).

The aim of the Informed Healthcare Choices (IHC) project is to develop and evaluate learning resources to help people to assess treatment claims and make informed healthcare 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, and the effects of the resources will be tested in randomized trials in Uganda (16, 17).

Until recently, there has not been a list of the Key Concepts that people must be able to understand and apply in order to assess claims about the effects of treatments. The development of such a list was the starting point for developing the IHC resources and the Claim Evaluation Tools, which will be used to evaluate the effects of the resources (18). Although a large number of studies have been conducted in different areas to improve critical thinking skills related to one or more of these Key Concepts, this research is heterogeneous and outcomes are measured inconsistently (19). Furthermore, available instruments used to map or evaluate people’s understanding of claims about treatment effects include only a handful of the Key Concepts (19). Given this diverse, international interest, a set of tools is needed both to evaluate the effects of interventions to promote understanding of the Key Concepts and to assess the extent to which people are able to assess claims about treatments and make informed healthcare choices.

The Claim Evaluation Tools consist of a battery of open-access, objective multiple-choice items addressing each of the Key Concepts. Researchers, teachers and others can select those that are relevant for specific populations or purposes. The items includes 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 (20). In another paper we have described the iterative development of the Claim Evaluation Tools, including qualitative and quantitative feedback from experts and end-users in Uganda, Kenya, Rwanda, Norway, United Kingdom and Australia (20). Based on suggested revisions by experts and end-users, we tailored the item content and formats with the intention of improving the reliability and validity of assessments. The items were found to have face validity by people with expertise in methodology, and end-users judged the items to be relevant and acceptable in their settings (see figure 1 for example of format). The Claim Evaluation Tools were developed in English, but is currently being translated into Lugandan (Uganda), Norwegian, German, Spanish (Mexico), and Chinese. This paper describes the psychometric testing including Rasch analysis of a sub-set of the Claim Evaluation Tools administered in English in English speaking populations in Uganda and Norway.

Please enter figure 1. Example of format

Rasch analysis relies on item response theory and is based on the requirement that the comparison of two people are independent of which items are used within the set of items assessing the same variable (21-24). Furthermore, in the Rasch model the total score across items completely describes a person's standing on a variable. Rasch analysis has been used successfully in many disciplines including health research, and can be applied to both dichotomous and polytomous data (21, 25, 26). Rasch analysis is used to check the degree to which this scoring and summing-up across items is defensible in the data collected (21, 23, 25). For example, if you want to compare the ability to assess claims in two groups, such as patients and health professionals, it is important to be sure that the items work in the same way in the two groups. Rasch analysis also provides an excellent basis for revising individual items. When checking the items' fit to the Rasch model, misfit to the model can be easily diagnosed and items can be revised to improve fit (27). In this way, the Rasch analysis represents a dynamic approach to achieving construct validity, in which revisions are informed by the evidence (27).

Objective

To describe the psychometric testing and Rasch analysis of the Claim Evaluation Tools in Uganda and Norway.

Methods

Rasch analysis and item response theory

Item response theory focuses on individual item responses and observed "trait" score estimates (28). A trait refers to a person attribute. In this paper, a person trait score estimate or person location estimate β refers to the attribute "ability to critically assess claims about treatment effects" and an item difficulty or item location estimate δ refers to an item's difficulty.

The item response theory model referred to as the simple logistic Rasch model for dichotomous responses is represented mathematically as P(x=1)=exp(β-δ)/(1+exp(β-δ)). It models the probability P of a correct answer (x=1) as the difference between the item difficulty and the person ability β-δ expressed in logits. The unit "logit" or "log-odds unit" refers to ln[(probability of success)/(probability of failure)]=β-δ (24). When the item difficulty matches the person ability, β=δ and the probability of a correct answer is 0.5 or 50%. When β>δ or β<δ the probability of answering correctly is above or below 50%, respectively.

When data conform to the Rasch model, ability is measured consistently across the trait level scale with low measurement error. Hence, reliability follows from the Rasch model. In item response theory, measurement error is calculated as the standard error of estimate (SE of estimate=1/√I) being the inverse of the square root of the test information function (I). SE of estimate varies across the trait level scale as we tend to be better at measuring persons where there is more information i.e., more items telling us the strengths (correct answers) and weaknesses (wrong answers) of a person. As most items' difficulty is located close to the middle of the person ability distribution we tend to measure less well at the extremes where there are fewer items and accordingly less information. The precision of measurement is quantified by standard errors in item response theory. Another benefit of item response theory measurement is that the scale of measurement is an interval logit scale.

As mentioned above, Rasch analysis is a very useful tool for diagnosing problems with the items. Revisions can then be made to improve fit, either by revising or deleting response-options with sub-optimal fit, or by deleting whole items, informed by the evidence (27). This can be done by visual inspection of fit to the Item Characteristic Curve, Bonferroni adjusted high chi-square probability, and exploring the mean z-score by response alternative. The Item Characteristic Curve indicates the Rasch model's theoretically expected probability of answering correctly as a function of ability on the latent trait scale (see Figure 2 for an example of an Item Characteristic Curve).

Please enter figure 2. The Item Characteristic Curve

The mean ability estimates of the four class intervals used to group the persons based on ability are marked on the scale (in logits) (in this example taken from one of the Claim Evaluation Tools' item sets: -1.83, -1.18, -0.70 and 0.50). The imposed observed values indicate the proportion of correct answer in each interval. The inflection point (0.52, 0.50) refers to the item difficulty estimate (0.52) and the probability of a correct answer (0.50) for the matching person ability. The dashed line is the upper asymptote for probability. See Tables 1 and 2 for examples of difficulty and fit statistics and item mean z-scores.

Please enter Table 1. Difficulty and fit statistics for the item

Please enter Table 2. Mean z-scores for the group choosing each response alternative

When the data conform to the Rasch model, the person's ability (what we measure) and the items (our measurement device) are independent. This is a necessary requirement of objective measurement referred to as *specific objectivity* (29). This item-invariant measurement of persons and person-invariant calibrations of items indicate *sample independence*. When the data conform to the Rasch model, the raw score sum of responses to the items (a measurement at the ordinal level) is a *sufficient statistic*; i.e., it contains all the information needed to estimate the person location. The Rasch model transforms ordinal data into an interval logit scale implying *additivity* on the scale (30-32). The requirement of invariance follows from the Rasch model (22). If an item difficulty measure is different for two levels of a person factor - that is, for two groups of respondents (for example children and adults) - that item exhibits "within item bias" referred to as DIF and invariance is violated (33).

The local independence assumption presupposes unidimensional data and no response dependence between pairs of items (24, 34). The requirement of unidimensionality is met when one latent trait sufficiently explains the variance shared among the item responses. It follows from this that one variable “explains” all the correlations between the items. If responding to the items requires abilities from two or more dimensions, two or more latent traits are needed to account for the correlations between the items. This indicates bi-dimensional or multidimensional data (i.e., a composite of several unidimensional subscales), and we would fit a multidimensional Rasch model using the framework of the Multidimensional Random Coefficients Multinomial Logit model to account for the correlations between the unidimensional subscales (35). The prescriptive Rasch model is confirmatory and could, in cases where the researchers are concerned about the dimensionality of their scale, be accompanied by confirmatory factor analysis (CFA). If an item offers clues to solve a subsequent item or pairs of items share specific similarities beyond what the latent trait(s) can account for, we refer to this as response dependence. The result is redundancy in the data and inefficient measures. Multidimensionality and response dependence violates the requirement of local independence.

Participants, setting and test administration

We included two samples in the psychometric testing of the items, a Ugandan sample including primary school children (approximately age >10) and adults, and a Norwegian sample including primary school children (approximately age >12). In both settings, the samples included people who had received some training in the Key Concepts through the pilots of the IHC learning resources. The Ugandan sample included health professionals. We recruited children, parents and other adults through our own networks and through networks of teachers and journalists that were established at the start of the IHC project (36, 37)

For this evaluation, we used classical psychometric testing for reliability (Cronbach's Alpha), for which a value of 0.7 or higher is considered adequate), and Rasch analysis. There is no consensus on the sample size needed to perform a Rasch analysis (38). This is a pragmatic judgement that takes account of the number of items being evaluated and the statistical power needed to identify DIF. We aimed to achieve approximately 250 respondents for each of four item sets. We also chose to include a group of children who had participated in a pilot of the IHC primary school resources at an international school in Norway. Although this was a small sample, it would give us some indication of the fit to the Rasch model in an international population, and provide information on difficulty and differential item functioning in the two different settings. In both settings, we administered the Claim Evaluation Tools in English, since this was the official school language in both the Norwegian International School as well as the Ugandan Schools.

We administrated all items addressing 22 of the 32 concepts (see Figure 1 for example of format). We judged ten of the 32 Key Concepts too difficult for people in our target groups in Uganda, and consequently items addressing these Key Concepts were not addressed in the IHC resources that were pilot tested (36, 37). Because of the large number of items to be tested, we distributed the items across four sets (of which the small Norwegian sample only responded to one set of items). Demographic variables included age (child/ adult), educational background (health professional or not health professional) and exposure to the concepts through IHC interventions (yes/ no). In the development of the Claim Evaluation Tools, low literacy skills were identified as a potential barrier in the Ugandan setting (20). Consequently, literacy was a variable that could introduce DIF in the Claim Evaluation Tool items data. To explore this, a four-item English reading test was included to explore text recognition and understanding as an indication of the respondents' literacy level (see Appendix 1). The items were designed to resemble the multiple-choice items addressing the Key Concepts. The first two items required the respondents to identify the correct text sequence in the scenario. The latter two items assessed whether the respondent was able to integrate the information in the scenario.

The components of the Rasch analysis

We scored all items dichotomously. Individual item fit statistics, such as Bonferroni adjusted chi-square probability, infit and t-value, were estimated using pairwise maximum likelihood estimation available in RUMM2030 and marginal maximum likelihood estimation available in ConQuest4 (39-45). Possible dimension violations and response violations of local independence were studied using the PCA/t-test procedure and by inspecting the residual correlation matrix estimated in RUMM2030 (46).

Dimensionality and response dependence might similarly be assessed using the unidimTest function in the ltm R-package and the Q3 statistic in the sirt R-package, respectively (47, 48). Our analysis of DIF for available person factors was carried out using RUMM2030.

Results

The total sample included 1114 people, among whom 685 were children and 429 were adults (including 171 health professionals). Of these 1114 people, 329 had received some form of training related to the Key Concepts. The Norwegian sample equalled 5% (59 respondents) out of the total respondents. The respondents completed most of the items, and the mean number of missing or incorrectly filled in responses was <1%. Less than 1/3 responded correctly to all four reading test questions in the Ugandan sample, indicating a low-literacy level in English in this particular sample.

Item discrimination and fit

Most of the items conformed well to the Rasch model and only a few items showed evidence of DIF. In set 1 (the only set applied in both Uganda and Norway), six items had DIF based on setting (Norway and Uganda). A few items in each of the four sets also had some DIF for age and group (see Table 3).

Please enter Table 3. Reliability, response dependency and DIF by item set

Targeting and reliability

Cronbach's Alpha was satisfactory with the exception of set 4, which was below the desired threshold (see Table 3). Overall, the items developed to assess claims about treatment effects had high difficulty in the target population, with no very easy items. However, there were no extremely difficult items. The observed logits for sets 1, 2, 3 and 4 were -0.81, -1.06, -1.15 and -1.15 respectively. The spread (standard deviation) in the variable «ability» was as expected.

Possible dimension and response violation of local independence

Results indicated that all four sets measured a sufficiently unidimensional latent trait. There were no specific sub-dimensions measuring different traits. Hence, one major dimension governs the responses to the items (use of principal component analysis of residuals/ dependent t-test procedure).

Only weak dependence was found in two items for set 1 with a residual correlation of 0.21. We did not observe any response dependence in the other three sets.

Discussion

We have developed the Claim Evaluation Tools using qualitative and quantitative feedback from methodologists and end-users in six countries. This study reports the findings of the first psychometric testing conducted in two settings, Uganda and Norway, of items addressing 22 out of 32 Key Concepts. Overall, we found that these subsets of the Claim Evaluation Tools had high reliability. Most of the items did not over-or underdiscriminate, or function differently across subgroups of participants. No important response dependence was identified, and all four sets measured a sufficiently unidimensional latent trait. The findings of this study will inform decisions about revising or deleting items with sub-optimal fit to the Rasch model.

Feedback on the items from experts and end-users suggested that the items were difficult (20). Our Rasch analysis confirmed this. Moreover, the respondents reading skills was found to be low in the target population in Uganda. This suggests that efforts should be made to simplify the text in the scenarios and editing the response options to improve readability. Removing response options in items with sub-optimal fit to the Rasch model could also contribute in making the items less difficult.

A limitation of this study is that we tested the items in only two settings, Uganda and Norway, and the fit with the Rasch model when used in other settings is unknown. Although we found few items with cross-cultural differential item functioning, any other application of the Claim Evaluation Tools in other countries should be tested using Rasch analysis. Furthermore, we did not include gender in the analysis, which could introduce DIF, and this will be explored in further testing.

There has been encouraging interest in using the Claim Evaluation Tools in settings other than the countries included in the IHC project, and researchers in Norway, Mexico, Germany and China are currently translating and testing the multiple-choice items in their settings. In addition, the items addressing the Key Concepts judged to be more advanced and which were not tested as part of this study are currently being tested online through testingtreatments.org including people with expertise in evidence based medicine. We are also developing items to assess intended behaviours, attitudes towards assessing treatment claims and self-efficacy.

All Claim Evaluation Tools will be freely available for non-commercial use on request through the Testing Treatments interactive website (www.testingtreatments.org). We will tag individual items by the settings in which the items have been tested, including information about the people in which the item was tested, and how it was tested (qualitative feedback or psychometric testing). We will also make the item's properties based on any psychometric testing available, including information about reliability and difficulty (22, 24). We will tag items that have been newly developed or revised as 'new' and 'awaiting assessment'. This will provide transparency and make the items easily accessible to those who would like to use them in teaching or for research activities.

For the pilots and statistical tests performed as part of this development work, we scored the items as percent correct per item and calculated the overall percent correct responses across items for individuals. However, mean scores can be difficult to interpret. To supplement this, it is possible to use an absolute (criterion referenced) standard to set a passing score i.e. a cut score; e.g. for passing or failing. The judgement about the required level of achievement is pragmatic, and there are several ways of doing this (49-51). For the items that will be used in the trial of the resources, we will establish criteria referenced standards using a combination of a combination of Nedelsky's and Angoff's methods (16, 17)(43-46).

Conclusion

Based on psychometric testing using Rasch analysis in Uganda and Norway, we found that the tested instruments were reliable and most of the items conformed to the Rasch model. Taken together with our previously reported findings, these results suggest that the Claim Evaluation Tools we have tested are reliable and have face and construct validity in the settings in which they have been tested. Items with sub-optimal fit will be considered deleted or repaired. We encourage further testing of the Claim Evaluation Tools in other contexts.

Authors' Contributions

AA, ØG, AO wrote the protocol and the IHC group provided comments to this protocol. The Claim Evaluation Tools were developed by the IHC group. AA coordinated all of the development and evaluation process with support from AO. DS, AN and KO performed the data collection and data entry from the field testing. ØG and AA prepared the data files for the analysis, and ØG conducted the Rasch analysis. AA authored this manuscript with significant input from the IHC group.

Acknowledgements

We are deeply grateful to all of the enthusiastic children, parents and teachers that contributed to this project. We would also like to thank the IHC advisory panel, and the other researchers and methodologists that provided their advice in the development process.

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.

Ethical approval

Ethical approval was sought by the IHC project representatives in each country.

Data sharing statement

All data are published as part of this study or can be found at datadryad.org. All Claim Evaluation Tools are available upon request for non-commercial use.

References

1. Lewis M, Orrock P, Myers S. Uncritical reverence in CM reporting: Assessing the scientific quality of Australian news media reports. Health Sociology Review. 2010;19(1):57-72.
2. Glenton C, Paulsen E, Oxman A. Portals to Wonderland? Health portals lead confusing information about the effects of health care. BMC Medical Informatics and Decision Making. 2005;5:7:8.
3. Moynihan R, Bero L, Ross-Degnan D, Henry D, Lee K, Watkins J, et al. Coverage by the news media of the benefits and risks of medications. The New England Journal of Medicine. 2000;342(22):1645-50.
4. Wolfe R, Sharp L, Lipsky M. Content and design attributes of antivaccination web sites. Journal of American Medical Association. 2002;287(24):3245-48.
5. Woloshin S, Schwartz L, Byram S, Sox H, Fischhoff B, Welch H. Women's understanding of the mammography screening debate. Archives of Internal Medicine 2000;160:1434-40.
6. Fox S, Duggan M. Health Online 20132013 09.04.2013. Available from: http://www.pewinternet.org/Reports/2013/Health-online.aspx.
7. Robinson E, Kerr C, Stevens A, Lilford R, Braunholtz D, Edwards S, et al. Lay public's understanding of equipoise and randomisation in randomised controlled trials. Research Support, Non-U.S. Gov't. NHS R&D HTA Programme, 2005 Mar. Report No.: 1366-5278 (Linking) Contract No.: 8.

8. Sillence E, Briggs P, Harris PR, Fishwick L. How do patients evaluate and make use of online health information? Social Science & Medicine. 2007;64(9):1853-62.
9. Horsley T, Hyde C, Santesso N, Parkes J, Milne R, Stewart R. Teaching critical appraisal skills in healthcare settings. Cochrane Database of Systematic Reviews. 2011(11).
10. Evans I, Thornton H, Chalmers I, P. G. Testing Treatments: better research for better healthcare. Second edition. London: Pinter & Martin Ltd2011. Available from: Available online at www.testingtreatments.org/new-edition/.
11. Chalmers I., Glasziou P., Badenoch D., Atkinson P., Austvoll-Dahlgren A., Oxman A. Evidence Live 2016: Promoting informed healthcare choices by helping people assess treatment claims. BMJ; 26.06.2016.
12. Taking shared decision making more seriously. Lancet. 2011;377(9768):784.
13. Shekelle PG, Pronovost PJ, Wachter RM, McDonald KM, Schoelles K, Dy SM, et al. The top patient safety strategies that can be encouraged for adoption now. Annals of Internal Medicine. 2013;158(5 Pt 2):365-8.
14. Stacey D, Légaré F, Col NF, Bennett CL, Barry MJ, Eden KB, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database of Systematic Reviews 2014, Issue 1 Art No: CD001431 DOI: 101002/14651858CD001431pub4.
15. Berkman N, Sheridan S, Donahue K, Halpern D, Crotty K. Low Health Literacy and Health Outcomes: An Updated Systematic Review. Annals of Internal Medicine. 2011;155(2):97-U89.
16. 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. Submitted manuscript. 2016.
17. 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 claims about the benefits and harms of treatments? Protocol for a randomized trial

Submitted manuscript. 2016.
18. 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. Journal of Evidence-Based Medicine. 2015;8(3):112-25.
19. Austvoll-Dahlgren A, Nsangi A, Semakula D. Key concepts people need to understand to assess claims about treatment effects: a systematic mapping review of interventions and evaluation tools. Submitted paper. 2016.
20. 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”. Submitted paper. 2016.
21. 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):1358-62.
22. Andrich D. Rasch models for measurement. Beverly Hills: Sage Publications; 1988.
23. Leonard M. Rasch Promises: a Layman's Guide to the Rasch Method of Item Analysis. Educational Research. 1980;22(3):188-92.
24. Rasch G. Probabilistic Models for Some Intelligence and Achivement Tests. Copenhagen: Danish Institue for Educational Research. Expanded Edition 1983: Chicago: MESA Press; 1960.
25. Guttersrud O, Dalane JO, Pettersen S. Improving measurement in nutrition literacy research using Rasch modelling: examining construct validity of stage-specific 'critical nutrition literacy' scales. Public Health Nutr. 2014;17(4):877-83.

26. Conaghan PG, Emerton M, Tennant A. Internal construct validity of the Oxford Knee Scale: evidence from Rasch measurement. Arthritis Rheum. 2007;57(8):1363-7.
27. Rasch analysis. http://www.rasch-analysis.com/. Accessed 2016.
28. Hambleton RK, Swaminathan H, Rogers HJ. Chapter 2. concepts models and features. Fundamentals of Item Response Theory. Newbery Park: Sage publications1991. p. 7-31.
29. Stenner AJ. Specific objectivity - local and general. Rasch Measurement Transactions, 1994, 8:3 p.374. http://www.rasch.org/rmt/rmt83e.htm.
30. Andrich D. Distinctions between assumptions and requirements in measurement in the social sciences. In J.A. Keats et al. (Eds.), Mathematical and Theoretical Systems (pp 7–16), North Holland: Elsevier Science Publishers BV. 1989.
31. Salzberger T. Does the Rasch Model Convert an Ordinal Scale into an Interval Scale?

. Rasch Measurement Transactions, 2010, 24:2 p 1273-1275 http://wwwraschorg/rmt/rmt242ahtm. 2010.
32. Perline R, Wright B, Wainer H. The Rasch model as additive conjoint measurement. Applied Psychological Measurement, 3, 237-255. 1979.
33. Brodersen J, Meads D, Kreiner S, Thorsen H, Doward L, McKenna S. Methodological aspects of differential item functioning in the Rasch model. J Med Econ. 2007;10:309 – 24.
34. Marais I, Andrich D. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. J Appl Meas. 2008;9(3):200-15.
35. Adams RJ, Wilson M, Wang WC. The Multidimensional Random Coefficients Multinomial Logit Model. Applied Psychological Measurement, 21(1), 1-23. 1997.
36. Nsangi A, Semakula D, Oxman AD, Sewankambo NK. Teaching children in low-income countries to assess claims about treatment effects: prioritization of key concepts. Journal of Evidence-Based Medicine. 2015;8(4):173-80.
37. Semakula D, Nsangi A, Oxman AD, Sewankambo NK. Priority setting for resources to improve the understanding of information about claims of treatment effects in the mass media. Journal of Evidence-Based Medicine. 2015;8(2):84-90.
38. Linacre J. Sample size and item calibration stability. Rasch Measurement Transactions. 1994;7:328.
39. Bland JM, Altman DG. Multiple Significance Tests - the Bonferroni Method .10. British Medical Journal. 1995;310(6973):170-.
40. Glass GV, Stanly JC. Statistical methods in education and psychology, New Jersey: Prentice-Hall. . 1970.
41. Linacre J. What do Infit and Outfit, Mean-square and Standardized mean? Rasch Measurement Transactions. 2002;16:2 p.878. http://www.rasch.org/rmt/rmt162f.htm.
42. Andrich D, Lou G. Conditional pairwise estimation in the Rach model for ordered categories using principal components. Journal of applied measurement, 4(3), 205–221. 2003.
43. Andrich D, Lyne A, Sheridan B, Luo G. RUMM2030: Rasch Unidimensional Measurement Model software [computer program]. Perth: RUMM Laboratory. 2009.
44. Wu ML, Adams RJ, Wilson MR, Haldane SA. ACER ConQuest Version 2: Generalised item response modelling software [computer program]. Camberwell: Australian Council for Educational Research. 2007.
45. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46(4), 443-459. . 1981.

46. 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 Journal of Statistics. 2014

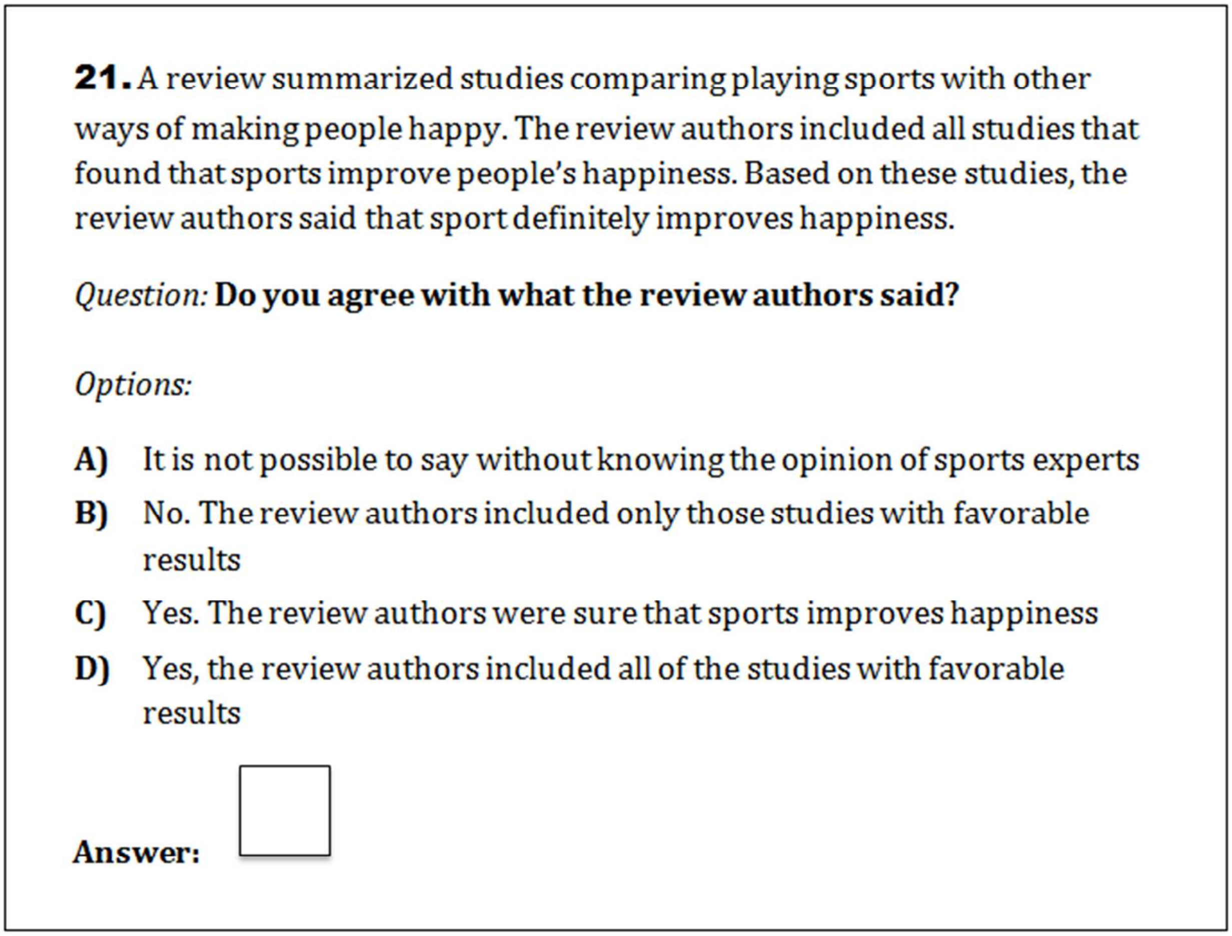
47. Robitzsch A. sirt: Supplementary Item Response Theory Models R package. http://CRAN.R-project.org/package=sirt. 2013.

48. Rizopoulos D. ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. Journal of Statistical Software, 17(5), 1-25 http://wwwjstatsoftorg/v17/i05/. 2006.

49. Livingston SA ZM. Passing scores; A manual for setting standards of performance on educational and occupational tests. Educational Testing Service. 1982.

50. Nedelsky L. Absolute grading standards for objective tests. Education and Psycholgical Measurement Journal. 1954;14(1):3-19.

51. Angoff WH. 4. Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL (ed.). Educational Measurement Washington DC. 1971:514-5.



21. A review summarized studies comparing playing sports with other ways of making people happy. The review authors included all studies that found that sports improve people's happiness. Based on these studies, the review authors said that sport definitely improves happiness.

Question: **Do you agree with what the review authors said?**

Options:

A) It is not possible to say without knowing the opinion of sports experts

B) No. The review authors included only those studies with favorable results

C) Yes. The review authors were sure that sports improves happiness

D) Yes, the review authors included all of the studies with favorable results

Answer: □

Figure 1. Example of format

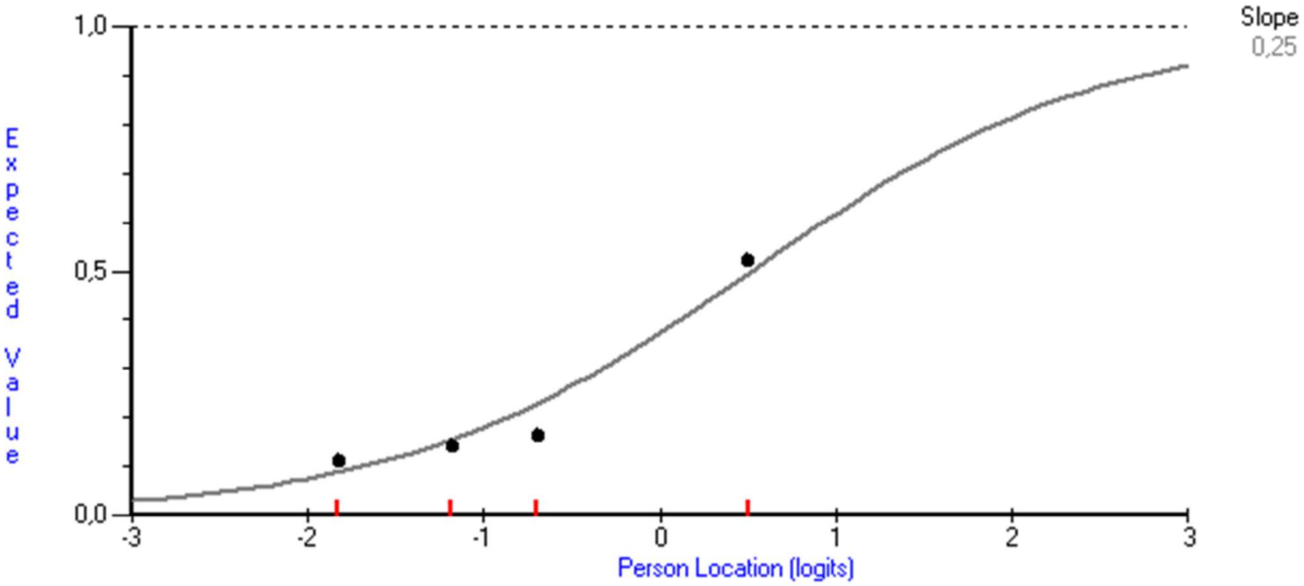


Figure 2. The Item Characteristic Curve

p	loc	SE	pb	z-fit	chi-sq	prob
0.25	0.52	0.16	0.46	-0,10	2.10	0.55

Table 1. Difficulty and fit statistics for the item shown in Figure 2 reported as (proportion of correct answers (p), item location parameter or difficulty (loc) estimated standard error (SE) of the item location parameter, item point biserial coefficient (pb), z-fit residual (z-fit), chi-square value (chi-sq estimated for n=255), chi-square probability

qn21	Mean z	p-value
1	-0.18	0.25
2*	**0.82**	**0.25**
3	-0.47	0.33
4	0.02	0.16
9	-0.50	0.02
Total	0.00	1.00

*indicates correct response

Table 2. Mean z-scores for the group choosing each response alternative

Set	n	Cronbach's alpha	Response dependence	DIF Age	DIF Group (Not in IHC pilots)	DIF Country
1	255	0.81	One pair of items (weak dependence)	0 items	1 item	6 items
2	287	0.78	Not observed	2 items	2 items	NR
3	289	0.78	Not observed	2 items	2 items	NR
4	283	0.63	Not observed	2 items	1 item	NR

Table 3. Reliability, response dependency and DIF by item set

1. A doctor did a study to find out if drinking tea keeps people from getting sick. The doctor tossed a coin to decide who should get the tea and who should not. People who got tea went to the doctor's office every day to drink their tea. At the end of the study, people who got the tea were less likely to be sick than those who got no tea.

Based on the text above, please answer the following questions:

1.1 Who went to the doctor's office every day?

Options:

A) People who did not get tea

B) People who got tea

C) Everyone

D) People who got sick

Answer: □

1.2. How did the doctor decide who should get tea?

Options:

A) By tossing a coin

B) By asking people what they would like

C) They gave tea to those who were more likely to be sick

D) They asked people who came to their office

Answer: □

1.3. What was the treatment?

Options:

A) Tea

B) Sleep

C) The study

D) The doctors

Answer: □

1.4. What was the result of the study?

Options:

A) Drinking tea can help people from getting sick

B) Doctors should toss coins when doing studies

C) People should go to the doctor if they are sick

D) Not drinking tea can help people from getting sick

Answer: □

BMJ Open

Measuring ability to assess claims about treatment effects: A latent trait analysis of items from the “Claim Evaluation Tools” database using Rasch modelling

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2016-013185.R1
Article Type:	Research
Date Submitted by the Author:	02-Feb-2017
Complete List of Authors:	Austvoll-Dahlgren, Astrid; Norwegian Institue of Public Health Guttersrud, Øystein; Norwegian Centre for Science Education, University of Oslo Nsangi, Allen; Makerere University College of Health Sciences, School of Medicine Semakula, Daniel; Makerere University College of Health Sciences, Oxman, Andrew; Norwegian Institue of Public Health
<b>Primary Subject Heading</b>:	Patient-centred medicine
Secondary Subject Heading:	Research methods, Public health, Health policy
Keywords:	evidence based medicine, health literacy, outcome measurement, multiple-choice, Rasch analysis, patient education

Measuring ability to assess claims about treatment effects: A latent trait analysis of items from the "Claim Evaluation Tools" database using Rasch modelling

Astrid Austvoll-Dahlgren, Øystein Guttersrud, Allen Nsangi, Daniel Semakula, Andrew D. Oxman, The IHC group*

Iain Chalmers

Leila Cusack

Claire Glenton

Tammy Hoffmann

Margaret Kaseje

Simon Lewin

Leah Atieno Marende

Michael Mugisha

Laetitia Nyirazinyoye

Kjetil Olsen

Matthew Oxman

Sarah Rosenbaum

Nelson K. Sewankambo

Anne Marie Uwitonze

Astrid Austvoll-Dahlgren (corresponding author)

astrid.austvoll-dahlgren@fhi.no

+47 41294057

Norwegian Institute of Public Health

BOKS 7004 St.Olavsplass

0130 Oslo, Norway

Øystein Guttersrud

oystein.guttersrud@naturfagsenteret.no

Norwegian Centre for Science Education, University of Oslo

Postboks 1106, Blindern 0317 Oslo, Norway

Allen Nsangi

nsallen2000@yahoo.com

Makerere University College of Health Sciences.

New Mulago Hospital Complex, Administration Building, Second Floor.

P.O.Box 7072, Kampala Uganda

Daniel Semakula

semakuladaniel@gmail.com

Makerere University College of Health Sciences.

New Mulago Hospital Complex, Administration Building, Second Floor.

P.O.Box 7072, Kampala Uganda

Andrew D. Oxman

oxman@online.no

Norwegian Institute of Public Health

BOKS 7004 St.Olavsplass

0130 Oslo, Norway

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

Abstract

Background: The Claim Evaluation Tools database contains multiple-choice items for measuring people's ability to apply the key concepts they need to know to be able to assess treatment claims. We assessed items from the database using Rasch analysis to develop an outcome measure to be used in two randomised trials in Uganda. Rasch analysis is a form of psychometric testing relying on Item Response Theory. It is a dynamic way of developing outcome measures that are valid and reliable.

Objectives: To assess the validity, reliability, and responsiveness of 88 items addressing 22 key concepts using Rasch analysis.

Participants: We administrated four sets of multiple-choice items in English to 1114 people in Uganda and Norway, of which 685 were children and 429 were adults (including 171 health professionals). We scored all items dichotomously. We explored summary and individual fit statistics using the RUMM2030 analysis package. We used SPSS to perform distractor analysis.

Results: Most items conformed well to the Rasch model but some items showed evidence of differential item functioning or needed revision. Overall, the four item sets had satisfactory reliability. We did not identify significant response dependence between any pairs of items and, overall, the magnitude of multidimensionality in the data was acceptable. The items had a high level of difficulty.

Conclusion: We found that most items that we tested had satisfactory fit to the Rasch model. Following revision of some items, informed by the findings from this study, we concluded that most of the items were suitable for use in an outcome measure for evaluating the ability of children or adults to assess treatment claims.

Strengths and limitations of this study

- To our knowledge, this is the first Rasch analysis of multiple-choice items that measure people's ability to assess claims about treatment effects.
- We have used robust methods to evaluated the items' validity and reliability in two settings, allowing for evidence informed revisions.
- Our analysis suggests that most items have acceptable model fit and can be used in the settings where they were tested.
- The items might function differently when translated or used in other settings.

Background

People are confronted with claims about treatment effects daily. This includes claims about the effects of changes in health behaviour, screening, other preventive interventions, therapeutic interventions, rehabilitation, and public health and health system interventions that are targeted at groups of people. A “treatment claim” is something someone says about whether a treatment causes something to happen or to change; for example, that Vitamin C prevents you from getting the common cold. A claim can be true or can be false (1-4). Many of these claims are not based on evidence from fair comparisons of treatments, and many patients and health professionals alike do not have the necessary skills to assess the reliability of these claims (5-11). Being able to think critically and make informed decisions is essential for engaging patients in clinical decisions and citizens in policy decisions (10, 12-14).

Interest in promoting critical thinking cuts across disciplines (15). There are many definitions and conceptualisations of critical thinking. In the learning sciences, critical thinking is defined as “purposeful, self-regulatory judgment that results in interpretation, analysis, evaluation, and inference, as well as explanations of the considerations on which that judgment is based” (16). There is debate about the extent to which critical thinking skills are “generic” and the extent to which they are content specific. Critical thinking is also a component of health literacy (17). In health literacy studies, critical thinking is content specific, focussing on people’s ability to think critically about health information. However, definitions of this component of health literacy are often fuzzy. They seldom describe which criteria patients should apply when thinking critically about health information (15). Critical thinking is also a key component of evidence-based practice. As in health literacy studies, critical thinking in evidence-based practice is content specific, but is operationalised as practical skills such as the ability to formulate questions, find relevant research, and assess the certainty of research evidence using explicit criteria (15, 18, 19).

Efforts to promote critical thinking as a component of evidence-based practice have largely focussed on health professionals. However, interest in helping patients and the public to make evidence-informed decisions is growing (11). One such initiative is the Informed Healthcare Choices (IHC) project, which aims to help people to 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 (20, 21).

The Claim Evaluation Tools database

The first step in the IHC project was to identify the Key Concepts people need to know to be able to assess treatment effects (22). This resulted in an initial list of 32 Key Concepts that serves as a syllabus for designing learning resources (22). This was also the starting point for the IHC learning resources. We present a short list of the Key Concepts in table 1. This list is hosted by testingtreatments.org, and is an evolving document subject to annual revisions.

Looking for suitable measurement tools to be used in the IHC trials, we conducted a systematic mapping review of interventions and assessment tools addressing the Key Concepts (15). Based on the findings of this review, we concluded that this research is heterogeneous and that outcomes are measured inconsistently (15). Furthermore, we found no instrument that addressed all the Key Concepts, or that would be suitable as an outcome measure in trials of the IHC learning resources. We therefore developed a database of multiple-choice items that could be used outcome measures in the two IHC trials in Uganda, as well for other purposes. The Claim Evaluation Tools database includes four or more items that address each of the Key Concepts. We developed these items in four steps, using qualitative and quantitative methods, over a three year period (2013-2016) (23):

(i) Determination of the scope of the database, writing and revising items;

(ii) Expert item review and feedback (face validity);

(iii) Cognitive interviews with end-users - including children, parents, teachers and patient representatives - to assess relevance, understanding and acceptability; and

(iv) Piloting and practical administrative tests of the items in different contexts.

Instead of a standard, fixed questionnaire, we wanted to create a database from which teachers and researchers can choose items relevant to their purposes and target groups, and design their own tests or questionnaires (23). We developed all items in English, but translations are now also available in Luganda (Uganda), Norwegian, German, Spanish (Mexico), and Chinese. Currently, the database includes approximately 190 items. The items are designed to be relevant across different contexts, and can be used for children (from ages 10 and up) and adults (including both patients and health professionals) (23). We use “one-best answer” response options in all items, with one answer being unambiguously the “best” and the remaining options “worse” (see figure 1 for an example of a multiple-choice item) (24).

We describe the development of the items in more detail elsewhere (23). We describe here the first psychometric testing, using Rasch analysis, of items from the Claim Evaluation Tools database. The items were selected for use in an outcome measure for trials of the IHC primary school resources and podcast. The purpose of the Rasch analysis is to ensure the validity and reliability of the outcome measure (25).

Table 1. Key Concepts that people need to understand to assess claims about treatment effects

Informed Health Choices Concepts
1. Recognising the need for fair comparisons of treatments *[Fair treatment comparisons are needed]*
1.1 Treatments may be harmful *[Treatments can harm]*
1.2 Personal experiences or anecdotes (stories) are an unreliable basis for determining the effects of most treatments *[Anecdotes are not reliable evidence]*
1.3 A treatment outcome may be associated with a treatment, but not caused by the treatment *[Association is not necessarily causation]*
1.4 Widely used or traditional treatments are not necessarily beneficial or safe [Practice is often not based on evidence]
1.5 New, brand-named, or more expensive treatments may not be better than available alternatives *[New treatments are not always better]*
1.6 Opinions of experts or authorities do not alone provide a reliable basis for deciding on the benefits and harms of treatments *[Expert opinion is not always right]*
1.7 Conflicting interests may result in misleading claims about the effects of treatments *[Be aware of conflicts of interest]*
1.8 Increasing the amount of a treatment does not necessarily increase the benefits of a treatment and may cause harm *[More is not necessarily better]*

1.9 Earlier detection of disease is not necessarily better *[Earlier is not necessarily better]*
1.10 Hope can lead to unrealistic expectations about the effects of treatments *[Avoid unrealistic expectations]*
1.11 Beliefs about how treatments work are not reliable predictors of the actual effects of treatments *[Theories about treatment can be wrong]*
1.12 Large, dramatic effects of treatments are rare *[Dramatic treatment effects are rare]*
2. Judging whether a comparison of treatments is a fair comparison *[Treatment comparisons should be fair]*
2.1 Evaluating the effects of treatments requires appropriate comparisons *[Treatment comparisons are necessary]*
2.2 Apart from the treatments being compared, the comparison groups need to be similar (i.e. 'like needs to be compared with like') *[Compare like with like]*
2.3 People’s experiences should be counted in the group to which they were allocated *[Base analyses on allocated treatment]*
2.4 People in the groups being compared need to be cared for similarly (apart from the treatments being compared) *[Treat comparison groups similarly]*
2.5 If possible, people should not know which of the treatments being compared they are receiving *[Blind participants to their treatments]*
2.6 Outcomes should be measured in the same way (fairly) in the treatment groups being compared *[Assess outcome measures fairly]*
2.7 It is important to measure outcomes in everyone who was included in the treatment comparison groups *[Follow up everyone included]*
3. Understanding the role of chance *[Understand the role of chance]*

3.1 Small studies in which few outcome events occur are usually not informative and the results may be misleading *[Small studies may be misleading]*
3.2 The use of p-values to indicate the probability of something having occurred by chance may be misleading; confidence intervals are more informative *[P-values alone can be misleading]*
3.3 Saying that a difference is statistically significant or that it is not statistically significant can be misleading *['Significance' may be misleading]*
4. Considering all of the relevant fair comparisons *[Consider all the relevant evidence]*
4.1 The results of single tests of treatments can be misleading *[Single studies can be misleading]*
4.2 Reviews of treatment tests that do not use systematic methods can be misleading *[Unsystematic reviews can mislead]*
4.3 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 *[Consider how certain the evidence is]*
5. Understanding the results of fair comparisons of treatments *[Understand the results of comparisons]*
5.1 Treatments may have beneficial and harmful effects *[Weigh benefits and harms of treatment]*
5.2 Relative effects of treatments alone can be misleading *[Relative effects can be misleading]*
5.3 Average differences between treatments can be misleading *[Average differences can be misleading]*
6. Judging whether fair comparisons of treatments are relevant *[Judge relevance of fair comparisons]*

6.1 Fair comparisons of treatments should measure outcomes that are important *[Outcomes studied may not be relevant]*
6.2 Fair comparisons of treatments in animals or highly selected groups of people may not be relevant *[People studied may not be relevant]*
6.3 The treatments evaluated in fair comparisons may not be relevant or applicable *[Treatments used may not be relevant]*
6.4 Results for a selected group of people within fair comparisons can be misleading *[Beware of subgroup analyses]*

Please enter figure 1. Example of a multiple choice-item taken from the Claim Evaluation Tools database

Objective

To assess the validity, reliability, and responsiveness of multiple-choice items from the Claim Evaluation Tools database, using Rasch analysis, in English speaking populations in Uganda and Norway.

Methods

Scope and setting

Most of the data collection took place in Uganda. The reason for this was that we intended to use the items from the Claim Evaluation Tools database as the primary outcome measure for the IHC trials there. The items in the Claim Evaluation tools database are expected to work in the same way for children and adults (23). Consequently, for this evaluation, we needed a sample including both children and adults to explore item bias (differential item functioning) associated with age. We also needed a mix of people with and without relevant training. For these purposes, we invited children (in year-5 of primary school, with a starting age of 10 years) and adults who had participated in piloting of the IHC resources. We also recruited children, parents and other adults (without training) through our networks in Uganda established at the start of the IHC project (26, 27).

We also included a group of children who had participated in a pilot of the IHC primary school resources at an international school in Norway. Although this was a small sample, it provided an indication of the fit to the Rasch model in an international population, and provide information on difficulty and differential item functioning in the two different settings.

Test administration and sample size

We evaluated 88 items addressing the 22 Key Concepts initially targeted by two IHC interventions (26, 27). Having multiple items for each concept allows us to delete items with poor fit to the Rasch model. In addition, Rasch analysis provides information on each item's difficulty. Having a range of items with different difficulties addressing each Key Concept can be useful for measurement purposes, for example when used in Computer Adaptive Testing.

There is no consensus on the sample size needed to perform a Rasch analysis (28). 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. Sine we intended to test many items, we did not consider it feasible to include these in a single test, and split the items into four sets or "tests". We aimed to include approximately 250 respondents in Uganda for each of the four tests.

The children in Norway only responded to one set (out of the four). In both settings, we administered the items in English, since this was the official school language in both the Norwegian international school and the Ugandan schools. The data collection took place in 2015.

In developing the Claim Evaluation Tools database, low literacy skills were identified as a potential barrier in the Ugandan setting (23). Consequently, we developed four items to evaluate the respondents' text recognition and understanding as an indication of their reading ability (see Appendix 1). We designed the items to resemble the multiple-choice items addressing the Key Concepts. The first two items required the respondents to identify the correct text in the scenario. The latter two items assessed whether the respondent understood the information in the scenario.

Rasch analysis

Rasch analysis is used to check the degree to which scoring and summing-up across items is defensible in the data collected (29, 30). 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)(30).

Rasch analysis has been used successfully in many disciplines including health research, and can be applied to both dichotomous and polytomous data (30-32). Rasch analysis also provides an excellent basis for developing and revising items, and in construction of item banks. Misfit to the Rasch model might be diagnosed and items can be deleted or revised to improve model fit (33). In this way, Rasch analysis represents a dynamic approach to achieving construct validity, in which revisions are informed by the evidence (33). For this analysis, we scored all items dichotomously. We used Excel for data entry, RUMM2030 for Rasch analysis and SPSS for a simple classical test theory approach to distractor analysis. We report the steps we took in our analysis below following the fundamental aspects of Rasch analysis (30).

Summary statistics and overall fit

In Rasch analysis, the response patterns to an item set are tested against what is expected by the model, i.e. the ratio between any two items should be constant across different ability groups (30). For this study, ability refers to the “ability to critically assess claims about treatment effects” In other words, the easier the item is, the more likely it will be answered correctly; and the more able the person is, the more likely he or she will answer correctly (34). We explored this relationship between the expected and observed data using the summary statistics function in RUMM2030.

The overall Item-Person Interaction is presented on a logit scale, where the mean item location is always given as a zero. The mean person location is relative to the mean item location, and a mean person location higher than and a mean person location higher or lower than “0” indicates that the test, on average, is “too easy” or “too hard”, respectively.

From this analysis, we also report the item and person Fit Residual Statistics; this assesses 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 (35). Ideally, item fit and person fit should have a mean of zero and a SD of one (30).

The Item-Trait- Interaction is an overall test of invariance to the set of items, and the extent to which the items are working as expected at the grouped ability levels (36). This is represented as a Chi-Square probability value. Ideally, this should be greater than 0.05, indicating that there is no “statistically significant” deviation between the observed and “expected” data.

Power of test of fit and reliability

The Person-Separation Index is an indicator of the power of a set of items to discriminate between ability groups (35). We considered a Person-Separation-Index greater than 0.7 to be acceptable (35). We also 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 solved this by calculating missing responses as “incorrect” responses (35).

Individual person and item fit

We investigated individual person fit. Misfit violates the principles of the Guttman structure and may indicate different types of error, such as guessing.

The Item Characteristic Curve indicates the Rasch model’s theoretically expected probability of answering correctly as a function of ability on the latent trait scale (see Figure 2 for an example of an Item Characteristic Curve). We inspected Item Characteristic Curves and used chi-square values as single item fit indices using a p = 5% as our significance level. Using Bonferroni adjustment, the significance level was adjusted (p = 0.05/k) according to the number of k significance tests carried out (one for each item).

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

Please insert figure 2. The Item Characteristic Curve

The curve in figure 2 represents the expected probability of answering correctly, and the dots represent the observed proportion of correct answers for some intervals of ability estimates (class intervals). When the observed proportions fit the curve, the data fit the Rasch model. Items with sub-optimal fit indicate measurement error (35).

When two groups of people (for example children and adults) respond differently to an item despite equal ability levels, that item displays “within item bias” or differential item functioning (DIF) and invariance is violated (37). We explored DIF using ANOVA in RUMM2030. There are two types of DIF. Uniform DIF is when one group of people perform consistently better on an item; for example, when an item is easier for all adults across all ability groups compared to children. This is less problematic than non-uniform DIF, where the differences between the groups vary across levels of the attribute (30).

Testing for multidimensionality and response dependency

Unidimensionality - having just one trait underlying respondents’ responses - is a fundamental requirement of measurement and is explored using Rasch analysis (34). Furthermore, there should be no response dependency in the data; i.e. people’s responses to an item should not have a bearing on their responses to other items (38). Response dependence results in redundancy in the data and inefficient measures.

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% (39). We also inspected the residual correlation matrix estimated in RUMM2030 (40). We considered residual correlations above 0.3 as indicators of response dependence between items (41).

Results

Description of sample

The total sample included 1114 people, among whom 685 were children and 429 were adults (including 171 health professionals). Of these 1114 people, 329 had received some form of training related to the Key Concepts. The Norwegian sample equalled 5% (59 respondents) out of the total respondents. The mean number of missing and incorrectly filled in responses was <1% per item set. Less than 1/3 responded correctly to all four reading test questions in the Ugandan sample, indicating a low literacy level in English.

Summary statistics and overall fit

Overall, the items were difficult with no very easy items and no extremely difficult items (see figure 2 for the Item Threshold distributions per set). The mean person locations per set were -0.81, -1.06, -1.15 and -1.15 logits, respectively. Fit Statistics are presented in table 2. Mean item fit residuals and person fit residuals were satisfactory and close to 0, although the standard errors for set 1 and 2 is somewhat higher than what we would like to see. The Item-Person maps for all sets are available in figures 3 to 6. The upper part of the Item-Person map represents respondents' ability levels; the lower part show the distribution of item locations. From this we can see that overall the items are difficult, suggesting that items with lower difficulties might be needed to make these tests more sensitive (able to discriminate between people at the lower end of the scale (those with lower ability) (35). The item-Trait interaction had a Chi square probability of 0.00 for all sets, indicating that not all items may work as expected.

The estimated reliability indices were also acceptable (Cronbach's Alpha > 0.70 for all sets with exception of set 4 where the value was 0.63). Similarly, the Person-Separation Indexes were satisfactory for all sets with the exception of set 4 which had a value of 0.54.

Table 2. Overall fit statistics and tests of local independence by set

Item set	**1**	**2**	**3**	**4**
Summary statistics				

Persons (n)	255	287	289	283
Mean item fit residual (sd)	0.02 (1.83)	-0.03 (1.52)	-0.14 (1.11)	-0.08 (1.38)
Mean person location (sd)	-0.81 (1.00)	-1.06 (0.97)	-1.15 (0.96)	-1.15 (0.75)
Overall chi-square value (df)	177.6 (75)	194.3 (96)	189.5(96)	170.7 (72)
Chi-square probability	0.00	0.00	0.00	0.00
Person Separation Index	0.75	0.70	0.69	0.54
Cronbach's Alpha	0.81	0.78	0.78	0.63
Items with misfit (n)	2	1	1	1
Persons with misfit (n)	0	2	1	3
Tests for multidimensionality and response dependency				
Proportion of significant paired t-tests (%)	3.5	2.8	2.8	5.7
Dependent pairs of items (n)	0	0	0	0

Please enter figure 3. Item Person map for item set 1

Please enter figure 4. Item Person map for item set 2.

Please enter figure 5. Item Person map for item set 3

Please enter figure 6. Item Person map for item set 4

Individual person and item fit

We identified few persons with misfit. Likewise, we identified only two items with potential misfit in set 1, and one item in the remaining three sets (see table 2).

Overall, most items fit well to the Rach model. However, out of the 88 items, seventeen items were found to have poor model fit. The findings resulting from the distractor analyses, also suggested that many of the items would be improved by deleting or revising response options.

In set 1 (the only set applied in both Uganda and Norway), five items displayed uniform DIF and one item displayed non-uniform DIF associated with setting (i.e., Norway and Uganda). Across the four sets, six items displayed uniform DIF associated with age.

Multidimensionality and response dependency

We did not observe any statistically significant residual correlation between pairs of items, and the paired *t*-test procedure indicated that the sets were sufficiently unidimensional (Table 2).

Discussion

We have developed the Claim Evaluation Tools database using qualitative and quantitative feedback from methodologists and end-users in six countries (23). This study reports the findings of the first psychometric testing of multiple-choice items from the database, using Rasch analysis conducted in two settings, Uganda and Norway.

Most of the 88 items (addressing 22 Key Concepts) conformed well to the Rasch model. However some items displayed DIF by setting and age, and some items required revisions. Overall, we found that the four item sets created from the Claim Evaluation Tools database had acceptable reliability. We did not identify significant response dependence between any pairs of items and, overall, the magnitude of multidimensionality in the data was acceptable.

Feedback on the items from experts and end-users suggested that the items were difficult (23). The Rasch analyses confirmed this. Moreover, the respondents reading skills were found to be low in the target population in Uganda. This suggests that efforts should be made to simplify the text in the scenarios and editing the response options to improve readability and improve validity. Reducing the number of response options in the items could also contribute to making the items less difficult.

Based on the findings from our analyses, we decided to remove items with non-uniform DIF from the Claim Evaluation Database. We also decided to revise items with poor model fit and reduce response options that did not work as expected. Revised items will be retested in the Ugandan context, where they will be used as an outcome measure.

A limitation of this study is that we tested the items in only two settings, Uganda and Norway, and that the fit to the Rasch model in other settings is unknown. Further testing of items from the Claim Evaluation Tools database using Rasch Analysis in other countries and languages is needed. We also did not include the respondents' gender in the analysis, which could introduce further DIF. This will be explored in further testing.

There has been an encouraging interest in the Claim Evaluation Tools database in settings other than the countries included in the IHC project, and researchers in Norway, Mexico, Germany and China are currently translating and testing the multiple-choice items in their settings. In addition, the items addressing the Key Concepts we judged to be more advanced, and which were not tested as part of this study, are currently being tested online through www.testingtreatments.org, targeting people with relevant training, such as health researchers or teachers of evidence-based medicine. We are also developing items to assess intended behaviours and attitudes towards assessing treatment claims. The Claim Evaluation Tools database, which includes all of these questions, as well as findings from evaluations such as this one, is freely available for non-commercial use on request through the Testing Treatments interactive website (www.testingtreatments.org).

When used for evaluating peoples' ability to assess treatment claims, an item set generated from the Claim Evaluation Tools database can be scored by calculating the number or percentage of correct responses. However, such scores can be difficult to interpret, especially when comparing the average score of two groups (e.g. in a randomised trial). An absolute (criterion referenced) standard for a passing score (i.e. a cut-off for passing) or for mastery of the Key Concepts that are tested. Setting a cut-off requires judgement, and there are several ways of doing this (42-44). For the items that will be used in the trials of the IHC primary school resources and podcast, we have established criteria referenced standards using a combination of Nedelsky's and Angoff's methods (42-45).

Conclusion

We found that most items that we tested had satisfactory fit to the Rasch model. Taken together with our previously reported findings, the findings of this study suggest that the items have face and construct validity in the settings in which they have been tested. Following revisions of some items, informed by the findings from this study, most of the items that we tested are suitable for use in an outcome measure that evaluates people's ability to apply the key concepts they need to know to be able to assess treatment claims.

Authors' contributions

AA, ØG, AO wrote the protocol and the IHC group provided comments to this protocol. The Claim Evaluation Tools database was developed by the IHC group. AA coordinated all of the development and evaluation process with support from AO. DS, AN and KO performed the data collection and data entry from the field-testing. ØG and AA prepared the data files for the analysis, and ØG conducted the Rasch analyses. AA authored this manuscript with significant input from the IHC group.

Acknowledgements

We are deeply grateful to all of the enthusiastic children, parents and teachers that contributed to this project. We would also like to thank the IHC advisory panel, and the other researchers and methodologists that provided their advice in the development process.

Funding and competing interests

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

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.

Data sharing statement

All data are published as part of this study; additional information is available upon request and on our website informedhealthchoices.org. All items in the Claim Evaluation Tools Database are available upon request for non-commercial use.

References

1. Lewis M, Orrock P, Myers S. Uncritical reverence in CM reporting: Assessing the scientific quality of Australian news media reports. Health Sociology Review. 2010;19(1):57-72.
2. Glenton C, Paulsen E, Oxman A. Portals to Wonderland? Health portals lead confusing information about the effects of health care. BMC Medical Informatics and Decision Making. 2005;5:7:8.
3. Moynihan R, Bero L, Ross-Degnan D, Henry D, Lee K, Watkins J, et al. Coverage by the news media of the benefits and risks of medications. The New England Journal of Medicine. 2000;342(22):1645-50.
4. Wolfe R, Sharp L, Lipsky M. Content and design attributes of antivaccination web sites. Journal of American Medical Association. 2002;287(24):3245-48.
5. Woloshin S, Schwartz L, Byram S, Sox H, Fischhoff B, Welch H. Women's understanding of the mammography screening debate. Archives of Internal Medicine 2000;160:1434-40.
6. Fox S, Duggan M. Health Online 20132013 09.04.2013. Available from: http://www.pewinternet.org/Reports/2013/Health-online.aspx.
7. Robinson E, Kerr C, Stevens A, Lilford R, Braunholtz D, Edwards S, et al. Lay public's understanding of equipoise and randomisation in randomised controlled trials. Research Support, Non-U.S. Gov't. NHS R&D HTA Programme, 2005 Mar. Report No.: 1366-5278 (Linking) Contract No.: 8.
8. Sillence E, Briggs P, Harris PR, Fishwick L. How do patients evaluate and make use of online health information? Social Science & Medicine. 2007;64(9):1853-62.
9. Horsley T, Hyde C, Santesso N, Parkes J, Milne R, Stewart R. Teaching critical appraisal skills in healthcare settings. Cochrane Database of Systematic Reviews. 2011(11).
10. Evans I, Thornton H, Chalmers I, P. G. Testing Treatments: better research for better healthcare. Second edition. London: Pinter & Martin Ltd2011. Available from: Available online at www.testingtreatments.org/new-edition/.
11. Chalmers I., Glasziou P., Badenoch D., Atkinson P., Austvoll-Dahlgren A., Oxman A. Evidence Live 2016: Promoting informed healthcare choices by helping people assess treatment claims. BMJ; 26.06.2016.
12. Taking shared decision making more seriously. Lancet. 2011;377(9768):784.
13. Stacey D, Légaré F, Col NF, Bennett CL, Barry MJ, Eden KB, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database of Systematic Reviews 2014, Issue 1 Art No: CD001431 DOI: 101002/14651858CD001431pub4.
14. Berkman N, Sheridan S, Donahue K, Halpern D, Crotty K. Low Health Literacy and Health Outcomes: An Updated Systematic Review. Annals of Internal Medicine. 2011;155(2):97-U89.
15. 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. Systematic Reviews 2016;5:215.
16. Abrami PC, Bernard RM, Borokhovski E, Waddington DI, Wade CA, T. P. Strategies for Teaching Students to Think Critically: A Meta-Analysis. Review of Educational Research June 2015, Vol 85, No 2, pp 275– 314.

17. Sorensen K, Van den Broucke S, Fullam J, Doyle G, Pelikan J, Slonska Z, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health. 2012;12:80.

18. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312(7023):71-2.

19. Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol. 2011;64(4):380-82.

20. 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. Submitted manuscript. 2016.

21. 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 claims about the benefits and harms of treatments? Protocol for a randomized trial

Submitted manuscript. 2016.

22. 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. Journal of Evidence-Based Medicine. 2015;8(3):112-25.

23. 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”. Accepted manuscript BMJ open. 2016.

24. Case SC, DB S. Constructing Written Test Questions For the Basic and Clinical Sciences (Third edition). Philadelphia, USA: 2002.

25. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol. 2007;6(12):1094-105.

26. Nsangi A, Semakula D, Oxman AD, Sewankambo NK. Teaching children in low-income countries to assess claims about treatment effects: prioritization of key concepts. Journal of Evidence-Based Medicine. 2015;8(4):173-80.

27. Semakula D, Nsangi A, Oxman AD, Sewankambo NK. Priority setting for resources to improve the understanding of information about claims of treatment effects in the mass media. Journal of Evidence-Based Medicine. 2015;8(2):84-90.

28. Linacre J. Sample size and item calibration stability. Rasch Measurement Transactions. 1994;7:328.

29. Leonard M. Rasch Promises: a Layman's Guide to the Rasch Method of Item Analysis. Educational Research. 1980;22(3):188-92.

30. 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):1358-62.

31. Guttersrud O, Dalane JO, Pettersen S. Improving measurement in nutrition literacy research using Rasch modelling: examining construct validity of stage-specific 'critical nutrition literacy' scales. Public Health Nutr. 2014;17(4):877-83.

32. Conaghan PG, Emerton M, Tennant A. Internal construct validity of the Oxford Knee Scale: evidence from Rasch measurement. Arthritis Rheum. 2007;57(8):1363-7.

33. Rasch analysis. http://www.rasch-analysis.com/. Accessed 2016.

34. 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.

35. Psylab Group. Introductory Rasch Analysis Using RUMM2030. The Section of Rehabilitation Medicine. University of Leeds: 2016.
36. Displaying the RUMM2030 Analysis. Rasch Unidimensional Measurement Model. 2015.
37. Brodersen J, Meads D, Kreiner S, Thorsen H, Doward L, McKenna S. Methodological aspects of differential item functioning in the Rasch model. J Med Econ. 2007;10:309 – 24.
38. Marais I, Andrich D. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. J Appl Meas. 2008;9(3):200-15.
39. RUMM. Extending the RUMM2030 Analysis. 7. ed: RUMM Laboratory Pty Ltd. 2009.
40. 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 Journal of Statistics. 2014
41. Andrich D., Humphry SM., Marais I. Quantifying local, response dependence between two polytomous items using the. Applied Psychological Measurement 36(4), 309–324. 2012.
42. Livingston SA ZM. Passing scores; A manual for setting standards of performance on educational and occupational tests. Educational Testing Service. 1982.
43. Nedelsky L. Absolute grading standards for objective tests. Education and Psycholgical Measurement Journal. 1954;14(1):3-19.
44. Angoff WH. 4. Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL (ed.). Educational Measurement Washington DC. 1971:514-5.
45. Davies A., Gerrity M., Nordheim L., Okebukola P., Opiyo N., Sharples J., et al. Measuring ability to assess claims about treatment effects: establishment of a standard for passing and mastery. IHC Working Paper 2017; ISBN 978-82-8082-802-6.

Figure legends

Figure 1. Example of a multiple choice-item taken from the Claim Evaluation Tools database

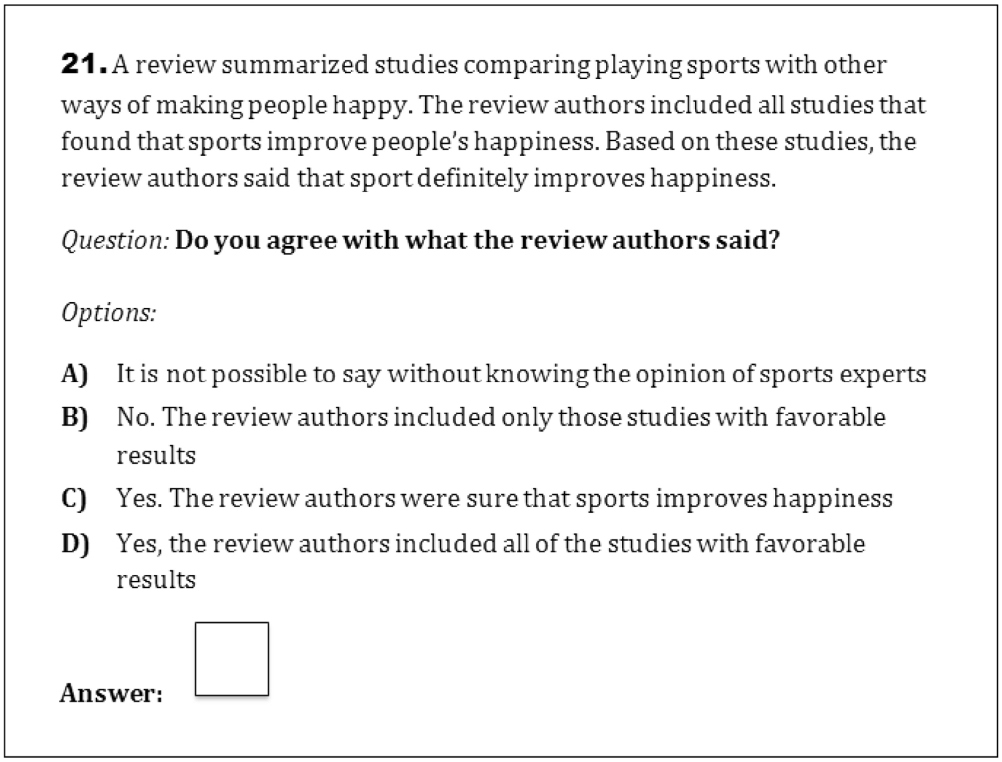
Figure 2. The Item Characteristic Curve

Figure 3. Item Person map for item set 1

Figure 4. Item Person map for item set 2.

Figure 5. Item Person map for item set 3

Figure 6. Item Person map for item set 4



21. A review summarized studies comparing playing sports with other ways of making people happy. The review authors included all studies that found that sports improve people's happiness. Based on these studies, the review authors said that sport definitely improves happiness.

Question: **Do you agree with what the review authors said?**

Options:

A) It is not possible to say without knowing the opinion of sports experts

B) No. The review authors included only those studies with favorable results

C) Yes. The review authors were sure that sports improves happiness

D) Yes, the review authors included all of the studies with favorable results

Answer: □

Figure 1. Example of a multiple choice-item taken from the Claim Evaluation Tools database

162x123mm (300 x 300 DPI)

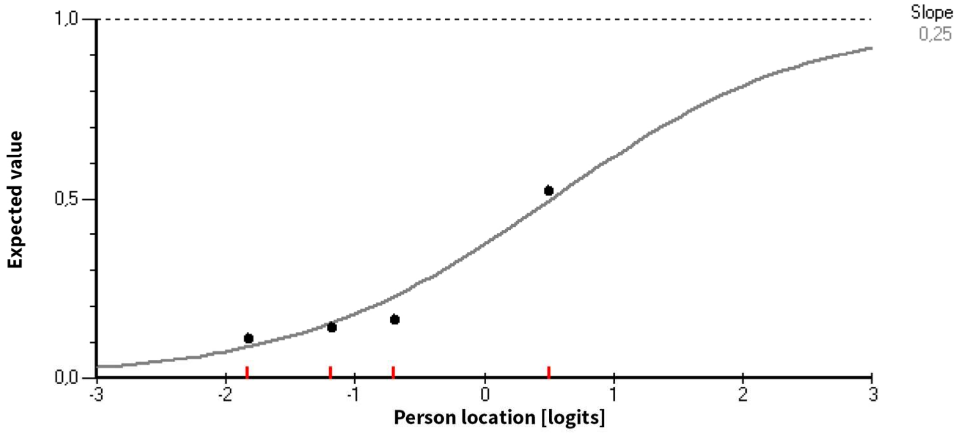


Figure 2. The Item Characteristic Curve

163x82mm (300 x 300 DPI)

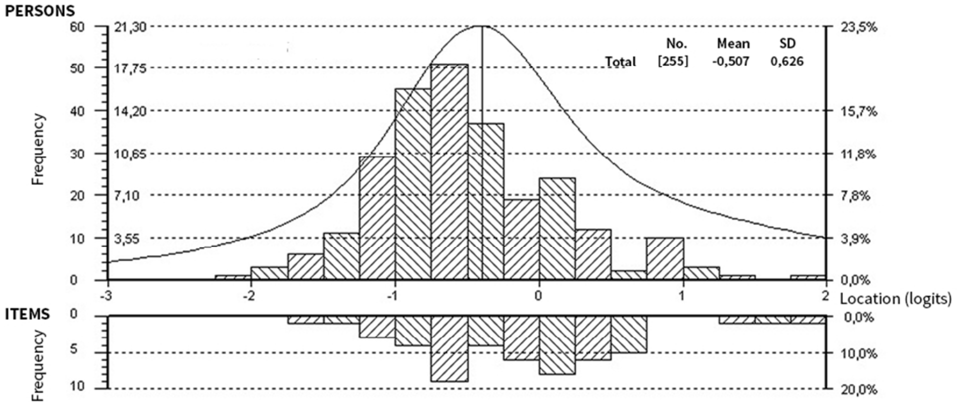


Figure 3. Item Person map for item set 1

82x37mm (300 x 300 DPI)

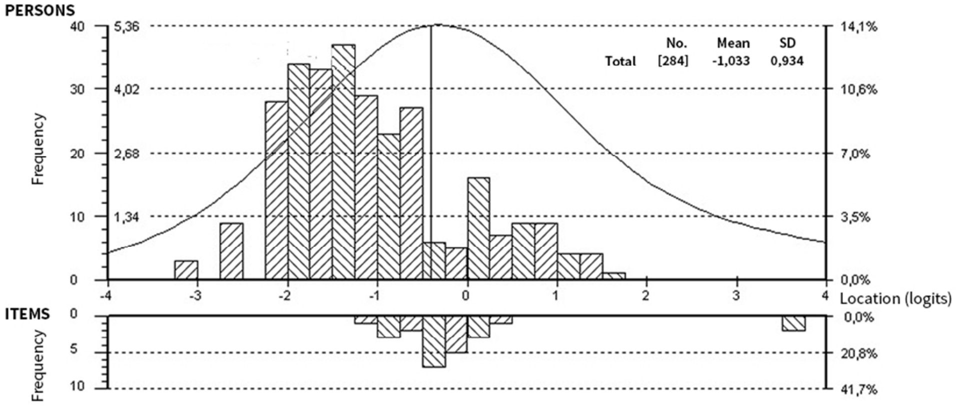


Figure 4. Item Person map for item set 2.

82x37mm (300 x 300 DPI)

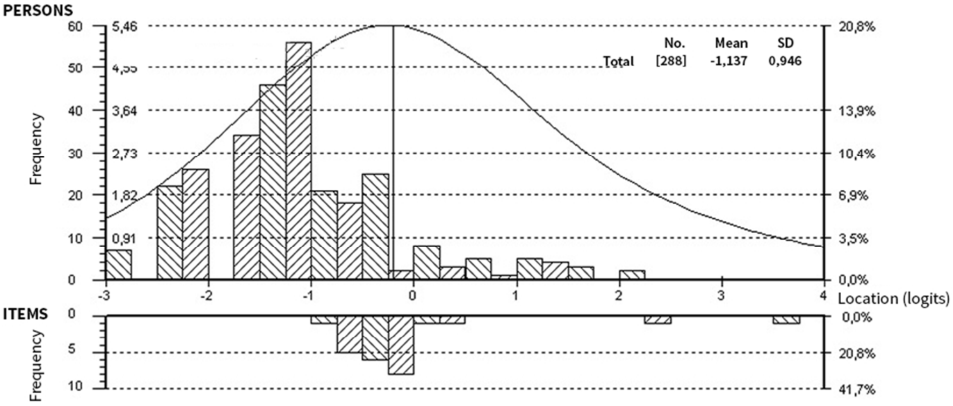


Figure 5. Item Person map for item set 3

82x37mm (300 x 300 DPI)

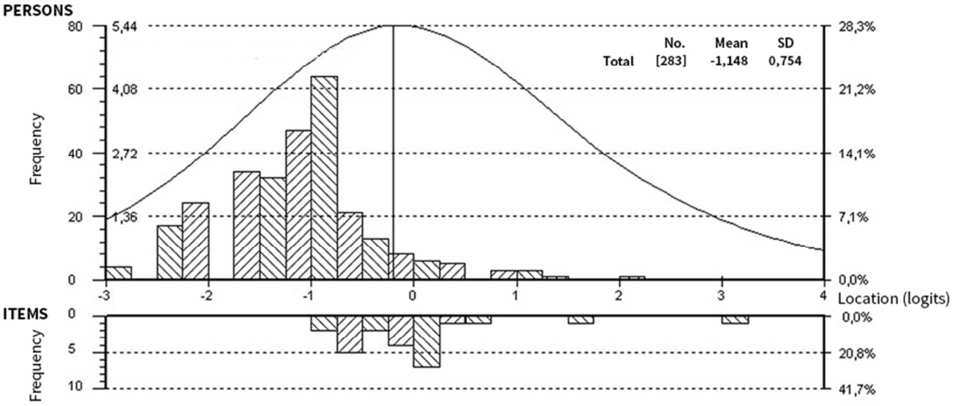


Figure 6. Item Person map for item set 4

82x37mm (300 x 300 DPI)

1. A doctor did a study to find out if drinking tea keeps people from getting sick. The doctor tossed a coin to decide who should get the tea and who should not. People who got tea went to the doctor's office every day to drink their tea. At the end of the study, people who got the tea were less likely to be sick than those who got no tea.

Based on the text above, please answer the following questions:

1.1 Who went to the doctor's office every day?

Options:

A) People who did not get tea

B) People who got tea

C) Everyone

D) People who got sick

Answer: □

1.2. How did the doctor decide who should get tea?

Options:

A) By tossing a coin

B) By asking people what they would like

C) They gave tea to those who were more likely to be sick

D) They asked people who came to their office

Answer: □

1.3. What was the treatment?

Options:

A) Tea

B) Sleep

C) The study

D) The doctors

Answer: □

1.4. What was the result of the study?

Options:

A) Drinking tea can help people from getting sick

B) Doctors should toss coins when doing studies

C) People should go to the doctor if they are sick

D) Not drinking tea can help people from getting sick

Answer: □

BMJ Open

Measuring ability to assess claims about treatment effects: A latent trait analysis of items from the “Claim Evaluation Tools” database using Rasch modelling

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2016-013185.R2
Article Type:	Research
Date Submitted by the Author:	10-Mar-2017
Complete List of Authors:	Austvoll-Dahlgren, Astrid; Norwegian Institue of Public Health Guttersrud, Øystein; Norwegian Centre for Science Education, University of Oslo Nsangi, Allen; Makerere University College of Health Sciences, School of Medicine Semakula, Daniel; Makerere University College of Health Sciences, Oxman, Andrew; Norwegian Institue of Public Health
<b>Primary Subject Heading</b>:	Patient-centred medicine
Secondary Subject Heading:	Research methods, Public health, Health policy
Keywords:	evidence based medicine, health literacy, outcome measurement, multiple-choice, Rasch analysis, patient education

Measuring ability to assess claims about treatment effects: A latent trait analysis of items from the "Claim Evaluation Tools" database using Rasch modelling

Astrid Austvoll-Dahlgren, Øystein Guttersrud, Allen Nsangi, Daniel Semakula, Andrew D. Oxman, The IHC group*

Iain Chalmers

Leila Cusack

Claire Glenton

Tammy Hoffmann

Margaret Kaseje

Simon Lewin

Leah Atieno Marende

Michael Mugisha

Laetitia Nyirazinyoye

Kjetil Olsen

Matthew Oxman

Sarah Rosenbaum

Nelson K. Sewankambo

Anne Marie Uwitonze

Astrid Austvoll-Dahlgren (corresponding author)

astrid.austvoll-dahlgren@fhi.no

+47 41294057

Norwegian Institute of Public Health

BOKS 7004 St.Olavsplass

0130 Oslo, Norway

Øystein Guttersrud

oystein.guttersrud@naturfagsenteret.no

Norwegian Centre for Science Education, University of Oslo

Postboks 1106, Blindern 0317 Oslo, Norway

Allen Nsangi

nsallen2000@yahoo.com

Makerere University College of Health Sciences.

New Mulago Hospital Complex, Administration Building, Second Floor.

P.O.Box 7072, Kampala Uganda

Daniel Semakula

semakuladaniel@gmail.com

Makerere University College of Health Sciences.

New Mulago Hospital Complex, Administration Building, Second Floor.

P.O.Box 7072, Kampala Uganda

Andrew D. Oxman

oxman@online.no

Norwegian Institute of Public Health

BOKS 7004 St.Olavsplass

0130 Oslo, Norway

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

Abstract

Background: The Claim Evaluation Tools database contains multiple-choice items for measuring people's ability to apply the key concepts they need to know to be able to assess treatment claims. We assessed items from the database using Rasch analysis to develop an outcome measure to be used in two randomised trials in Uganda. Rasch analysis is a form of psychometric testing relying on Item Response Theory. It is a dynamic way of developing outcome measures that are valid and reliable.

Objectives: To assess the validity, reliability, and responsiveness of 88 items addressing 22 key concepts using Rasch analysis.

Participants: We administrated four sets of multiple-choice items in English to 1114 people in Uganda and Norway, of which 685 were children and 429 were adults (including 171 health professionals). We scored all items dichotomously. We explored summary and individual fit statistics using the RUMM2030 analysis package. We used SPSS to perform distractor analysis.

Results: Most items conformed well to the Rasch model but some items needed revision. Overall, the four item sets had satisfactory reliability. We did not identify significant response dependence between any pairs of items and, overall, the magnitude of multidimensionality in the data was acceptable. The items had a high level of difficulty.

Conclusion: Most of the items conformed well to the Rasch model's expectations. Following revision of some items, we concluded that most of the items were suitable for use in an outcome measure for evaluating the ability of children or adults to assess treatment claims.

Strengths and limitations of this study

- To our knowledge, this is the first Rasch analysis of multiple-choice items that measure people's ability to assess claims about treatment effects.
- We have used robust methods to evaluate the items' validity and reliability in two settings, allowing for evidence informed revisions.
- Our analyses suggest that most items have acceptable model fit and can be used in the settings where they were tested.
- The items might function differently when translated or used in other settings.

Background

People are confronted with claims about treatment effects daily. This includes claims about the effects of changes in health behaviour, screening, other preventive interventions, therapeutic interventions, rehabilitation, and public health and health system interventions that are targeted at groups of people. A "treatment claim" is something someone says about whether a treatment causes something to happen or to change; for example, that Vitamin C prevents you from getting the common cold. A claim can be true or can be false (1-4). Many of these claims are not based on evidence from fair comparisons of treatments, and many patients and health professionals alike do not have the necessary skills to assess the reliability of these claims (5-11). Being able to think critically and make informed decisions is essential for engaging patients in clinical decisions and citizens in policy decisions (10, 12-14).

Interest in promoting critical thinking cuts across disciplines (15). There are many definitions and conceptualisations of critical thinking. In the learning sciences, critical thinking is defined as "purposeful, self-regulatory judgment that results in interpretation, analysis, evaluation, and inference, as well as explanations of the considerations on which that judgment is based" (16). There is debate about the extent to which critical thinking skills are "generic" and the extent to which they are content specific. Critical thinking is also a component of health literacy (17). In health literacy studies, critical thinking is content specific, focussing on people's ability to think critically about health information. However, definitions of this component of health literacy are often fuzzy. They seldom describe which criteria patients should apply when thinking critically about health information (15). Critical thinking is also a key component of evidence-based practice. As in health literacy studies, critical thinking in evidence-based practice is content specific, but is operationalised as practical skills such as the ability to formulate questions, find relevant research, and assess the certainty of research evidence using explicit criteria (15, 18, 19).

Efforts to promote critical thinking as a component of evidence-based practice have largely focussed on health professionals. However, interest in helping patients and the public to make evidence-informed decisions is growing (11). One such initiative is the Informed Healthcare Choices (IHC) project, which aims to help people to 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 (20, 21).

The Claim Evaluation Tools database

The first step in the IHC project was to identify the Key Concepts people need to know to be able to assess treatment effects (22). This resulted in an initial list of 32 Key Concepts that serves as a syllabus for designing learning resources (22). This was also the starting point for the IHC learning resources. We present a short list of the Key Concepts in table 1. This list is hosted by testingtreatments.org, and is an evolving document subject to annual revisions.

Looking for suitable measurement tools to be used in the IHC trials, we conducted a systematic mapping review of interventions and assessment tools addressing the Key Concepts (15). Based on the findings of this review, we concluded that this research is heterogeneous and that outcomes are measured inconsistently (15). Furthermore, we found no instrument that addressed all the Key Concepts, or that would be suitable as an outcome measure in trials of the IHC learning resources. We therefore developed a database of multiple-choice items that could be used outcome measures in the two IHC trials in Uganda, as well for other purposes. The Claim Evaluation Tools database includes four or more items that address each of the Key Concepts. We developed these items in four steps, using qualitative and quantitative methods, over a three year period (2013-2016) (23):

(i) Determination of the scope of the database, writing and revising items;
(ii) Expert item review and feedback (face validity);
(iii) Cognitive interviews with end-users - including children, parents, teachers and patient representatives - to assess relevance, understanding and acceptability; and
(iv) Piloting and practical administrative tests of the items in different contexts.

Instead of a standard, fixed questionnaire, we wanted to create a database from which teachers and researchers can choose items relevant to their purposes and target groups, and design their own tests or questionnaires (23). We developed all items in English, but translations are now also available in Luganda (Uganda), Norwegian, German, Spanish (Mexico), and Chinese. Currently, the database includes approximately 190 items. The items are designed to be relevant across different contexts, and can be used for children (from ages 10 and up) and adults (including both patients and health professionals) (23). We use "one-best answer" response options in all items, with one answer being unambiguously the “best” and the remaining options “worse” (see figure 1 for an example of a multiple-choice item) (24).

We describe the development of the items in more detail elsewhere (23). We describe here the first psychometric testing, using Rasch analysis, of items from the Claim Evaluation Tools database. The items were selected for use in an outcome measure for trials of the IHC primary school resources and podcast. The purpose of the Rasch analysis is to ensure the validity and reliability of the outcome measure (25).

Table 1. Key Concepts that people need to understand to assess claims about treatment effects

Informed Health Choices Concepts
1. Recognising the need for fair comparisons of treatments *[Fair treatment comparisons are needed]*
1.1 Treatments may be harmful *[Treatments can harm]*
1.2 Personal experiences or anecdotes (stories) are an unreliable basis for determining the effects of most treatments *[Anecdotes are not reliable evidence]*
1.3 A treatment outcome may be associated with a treatment, but not caused by the treatment *[Association is not necessarily causation]*
1.4 Widely used or traditional treatments are not necessarily beneficial or safe [Practice is often not based on evidence]
1.5 New, brand-named, or more expensive treatments may not be better than available alternatives *[New treatments are not always better]*
1.6 Opinions of experts or authorities do not alone provide a reliable basis for deciding on the benefits and harms of treatments *[Expert opinion is not always right]*
1.7 Conflicting interests may result in misleading claims about the effects of treatments *[Be aware of conflicts of interest]*
1.8 Increasing the amount of a treatment does not necessarily increase the benefits of a treatment and may cause harm *[More is not necessarily better]*

1.9 Earlier detection of disease is not necessarily better *[Earlier is not necessarily better]*
1.10 Hope can lead to unrealistic expectations about the effects of treatments *[Avoid unrealistic expectations]*
1.11 Beliefs about how treatments work are not reliable predictors of the actual effects of treatments *[Theories about treatment can be wrong]*
1.12 Large, dramatic effects of treatments are rare *[Dramatic treatment effects are rare]*
2. Judging whether a comparison of treatments is a fair comparison *[Treatment comparisons should be fair]*
2.1 Evaluating the effects of treatments requires appropriate comparisons *[Treatment comparisons are necessary]*
2.2 Apart from the treatments being compared, the comparison groups need to be similar (i.e. 'like needs to be compared with like') *[Compare like with like]*
2.3 People's experiences should be counted in the group to which they were allocated *[Base analyses on allocated treatment]*
2.4 People in the groups being compared need to be cared for similarly (apart from the treatments being compared) *[Treat comparison groups similarly]*
2.5 If possible, people should not know which of the treatments being compared they are receiving *[Blind participants to their treatments]*
2.6 Outcomes should be measured in the same way (fairly) in the treatment groups being compared *[Assess outcome measures fairly]*
2.7 It is important to measure outcomes in everyone who was included in the treatment comparison groups *[Follow up everyone included]*
3. Understanding the role of chance *[Understand the role of chance]*

3.1 Small studies in which few outcome events occur are usually not informative and the results may be misleading *[Small studies may be misleading]*
3.2 The use of p-values to indicate the probability of something having occurred by chance may be misleading; confidence intervals are more informative *[P-values alone can be misleading]*
3.3 Saying that a difference is statistically significant or that it is not statistically significant can be misleading *['Significance' may be misleading]*
4. Considering all of the relevant fair comparisons *[Consider all the relevant evidence]*
4.1 The results of single tests of treatments can be misleading *[Single studies can be misleading]*
4.2 Reviews of treatment tests that do not use systematic methods can be misleading *[Unsystematic reviews can mislead]*
4.3 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 *[Consider how certain the evidence is]*
5. Understanding the results of fair comparisons of treatments *[Understand the results of comparisons]*
5.1 Treatments may have beneficial and harmful effects *[Weigh benefits and harms of treatment]*
5.2 Relative effects of treatments alone can be misleading *[Relative effects can be misleading]*
5.3 Average differences between treatments can be misleading *[Average differences can be misleading]*
6. Judging whether fair comparisons of treatments are relevant *[Judge relevance of fair comparisons]*

6.1 Fair comparisons of treatments should measure outcomes that are important *[Outcomes studied may not be relevant]*
6.2 Fair comparisons of treatments in animals or highly selected groups of people may not be relevant *[People studied may not be relevant]*
6.3 The treatments evaluated in fair comparisons may not be relevant or applicable *[Treatments used may not be relevant]*
6.4 Results for a selected group of people within fair comparisons can be misleading *[Beware of subgroup analyses]*

Please enter figure 1. Example of a multiple choice-item taken from the Claim Evaluation Tools database

Objective

To assess the validity, reliability, and responsiveness of multiple-choice items from the Claim Evaluation Tools database, using Rasch analysis, in English speaking populations in Uganda and Norway.

Methods

Scope and setting

Most of the data collection took place in Uganda. The reason for this was that we intended to use the items from the Claim Evaluation Tools database as the primary outcome measure for the IHC trials there. The items in the Claim Evaluation tools database are expected to work in the same way for children and adults (23). Consequently, for this evaluation, we needed a sample including both children and adults to explore item bias (differential item functioning) associated with age. We also needed a mix of people with and without relevant training. For these purposes, we invited children (in year-5 of primary school, with a starting age of 10 years) and adults who had participated in piloting of the IHC resources. We also recruited children, parents and other adults (without training) through our networks in Uganda established at the start of the IHC project (26, 27).

We also included a group of children who had participated in a pilot of the IHC primary school resources at an international school in Norway. Although this was a small sample, it provided an indication of the fit to the Rasch model in an international population, and provide information on difficulty and differential item functioning in the two different settings.

Test administration and sample size

We evaluated 88 items addressing the 22 Key Concepts initially targeted by two IHC interventions (26, 27). Having multiple items for each concept allows us to delete items with poor fit to the Rasch model. In addition, Rasch analysis provides information on each item's difficulty. Having a range of items with different difficulties addressing each Key Concept can be useful for measurement purposes, for example when used in Computer Adaptive Testing.

There is no consensus on the sample size needed to perform a Rasch analysis (28). 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. Sine we intended to test many items, we did not consider it feasible to include these in a single test, and split the items into four sets or "tests". We aimed to include approximately 250 respondents in Uganda for each of the four tests.

The children in Norway only responded to one set (out of the four). In both settings, we administered the items in English, since this was the official school language in both the Norwegian international school and the Ugandan schools. The data collection took place in 2015.

In developing the Claim Evaluation Tools database, low literacy skills were identified as a potential barrier in the Ugandan setting (23). Consequently, we developed four items to evaluate the respondents' text recognition and understanding as an indication of their reading ability (see Appendix 1). We tested these items using the Lexile Analyser, and the items were found to fit within typical reading measures for 5th graders (29). We designed the items to resemble the multiple-choice items addressing the Key Concepts. The first two items required the respondents to identify the correct text in the scenario. The latter two items assessed whether the respondent understood the information in the scenario.

Rasch analysis

Rasch analysis is used to check the degree to which scoring and summing-up across items is defensible in the data collected (30, 31). 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 for multidimensionality, invariance of the items (item-person-interaction), and item bias (differential item function)(31).

Rasch analysis has been used successfully in many disciplines including health research, and can be applied to both dichotomous and polytomous data (31-33). Rasch analysis also provides an excellent basis for developing and revising items, and in construction of item banks. Misfit to the Rasch model might be diagnosed and items can be deleted or revised to improve model fit (34). In this way, Rasch analysis represents a dynamic approach to achieving construct validity, in which revisions are informed by the evidence (34). For this analysis, we scored all items dichotomously. We used Excel for data entry, RUMM2030 for Rasch analysis and SPSS for a simple classical test theory approach to distractor analysis. We report the steps we took in our analysis below following the fundamental aspects of Rasch analysis (31).

Summary statistics and overall fit

In Rasch analysis, the response patterns to an item set are tested against what is expected by the model, i.e. the ratio between any two items should be constant across different ability groups (31). For this study, ability refers to the “ability to critically assess claims about treatment effects”. In other words, the easier the item is, the more likely it will be answered correctly; and the more able the person is, the more likely he or she will answer correctly (35). We explored this relationship between the expected and observed data using the summary statistics function in RUMM2030.

The overall Item-Person Interaction is presented on a logit scale, and in RUMM2030 the mean item location is always given as a zero. A mean person location higher than “0” indicates that on average, the test is “too easy” and that the response group have a higher ability then the difficult level of the test. A mean person location lower than “0” suggests the test is “too hard”.

From this analysis, we also report the item and person Fit Residual Statistics; this assesses 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 (36). Ideally, item fit and person fit should have a mean of zero and a SD of one (31).

The Item-Trait- Interaction in RUMM2030 is a test of invariance to the scale, and whether or not the data fit the model for the discreet ability groups (37). Ideally, the Chi-Square probability value should be greater than 0.05, indicating that there is no “statistically significant” deviation between the observed data and what is expected form the model.

Power of test of fit and reliability

The Person-Separation Index is an indicator of the power of a set of items to discriminate between ability groups and individuals (36). We considered a Person-Separation-Index greater than 0.7 to be acceptable (36). We also 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. In RUMM2030, Cronbach’s Alpha can only be estimated if there are no missing data. We solved this by coding missing responses as “incorrect” responses (36).

Individual person and item fit

We investigated individual person fit. Misfit violates the principles of the Guttman structure and may indicate different types of error, such as guessing.

The Item Characteristic Curve indicates the theoretically expected probability of answering correctly as a function of ability on the latent trait scale (see Figure 2 for an example of an Item Characteristic Curve). We inspected Item Characteristic Curves and used Chi-square values as single item fit indices using a 0.05 as our significance level. Using Bonferroni adjustment, the significance level was adjusted (p = 0.05/k) according to the number of *k* significance tests carried out (one for each item).

We also performed distractor analysis using SPSS. This is particularly useful when developing and revising multiple-choice items, because it may identify response options that are not working as intended and can subsequently be deleted or revised.

Please insert figure 2. The Item Characteristic Curve

The curve in figure 2 represents the expected probability of answering correctly, and the dots represent the observed proportion of correct answers for some intervals of ability estimates (class intervals). When the observed proportions fit the curve, the data fit the Rasch model. Items with sub-optimal fit indicate measurement error (36).

When two groups of people (for example children and adults) respond differently to an item despite equal ability, that item displays “within item bias” or differential item functioning (DIF) and invariance is violated (38). There are two types of DIF. Uniform DIF is when one group of people perform consistently better on an item; for example, when an item is easier for all adults across all ability groups compared to children. This is less problematic than non-uniform DIF, where the differences between the groups vary across levels of the attribute (31).

We explored DIF for setting and age using ANOVA in RUMM2030. We also explored DIF by reading ability. This was done by pragmatically categorising the responses to the four reading ability items into two groups (merging respondents with 0, 1 or 2 correct responses and those with 3 or 4 correct responses).

Testing for multidimensionality and response dependency

Unidimensionality - having just one trait underlying responses - is a fundamental requirement of measurement and is explored using Rasch analysis (35). Furthermore, there should be no response dependency in the data; i.e. people’s responses to an item should not have a bearing on their responses to other items (39). Response dependence results in redundancy in the data and inefficient measures.

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% (40). We also inspected the residual correlation matrix estimated in RUMM2030 (41). We considered residual correlations above 0.3 as indicators of response dependence between items (42).

Results

Description of sample

The total sample included 1114 people, among whom 685 were children and 429 were adults (including 171 health professionals). Of these 1114 people, 329 had received some form of training related to the Key Concepts. The Norwegian sample equalled 5% (59 respondents) out of the total respondents. The mean number of missing and incorrectly filled in responses was <1% per item set. Less than 1/3 responded correctly to all four reading test questions in the Ugandan sample.

Summary statistics and overall fit

Overall, the items were difficult with no very easy items and no extremely difficult items (see figure 2 for the item threshold distributions per set). The mean person locations per set were -0.81, -1.06, -1.15 and -1.15 logits, respectively. Fit Statistics are presented in table 2. Mean item fit residuals and person fit residuals were satisfactory and close to 0, although the standard errors for set 1 and 2 is somewhat higher than what we would like to see. The Item-Person maps for all sets are available in figures 3 to 6. The upper part of the Item-Person map represents respondents' ability levels; the lower part show the distribution of item locations. From this we can see that, overall, the tests are difficult. This suggests that that easier items might be needed to make these tests more sensitive (able to separate between people at the lower end of the scale (those with lower ability) (36). The Chi square probability was 0.00 for all sets, indicating that not all items may work as expected.

The estimated reliability indices were acceptable (Cronbach's Alpha > 0.70 for all sets with exception of set 4 where the value was 0.63). Similarly, the Person-Separation Indexes were satisfactory for all sets with the exception of set 4 which had a value of 0.54.

Table 2. Overall fit statistics and tests of local independence by set

Item set	**1**	**2**	**3**	**4**
Summary statistics				
Persons (n)	255	287	289	283
Mean item fit residual (sd)	0.02 (1.83)	-0.03 (1.52)	-0.14 (1.11)	-0.08 (1.38)
Mean person location (sd)	-0.81 (1.00)	-1.06 (0.97)	-1.15 (0.96)	-1.15 (0.75)
Overall Chi-square value (df)	177.6 (75)	194.3 (96)	189.5(96)	170.7 (72)
Chi-square probability	0.00	0.00	0.00	0.00
Person Separation Index	0.75	0.70	0.69	0.54
Cronbach's Alpha	0.81	0.78	0.78	0.63
Items with misfit (n)	2	1	1	1
Persons with misfit (n)	0	2	1	3
Tests for multidimensionality and response dependency				
Proportion of significant paired t-tests (%)	3.5	2.8	2.8	5.7
Dependent pairs of items (n)	0	0	0	0

Please enter figure 3. Item Person map for item set 1

Please enter figure 4. Item Person map for item set 2.

Please enter figure 5. Item Person map for item set 3

Please enter figure 6. Item Person map for item set 4

Individual person and item fit

We identified few persons with misfit. Likewise, we identified only two items with potential misfit in set 1, and one item in the remaining three sets (see table 2).

Overall, most items fit well to the Rach model. However, out of the 88 items, seventeen items were found to have poor model fit. The findings resulting from the distractor analyses, also suggested that many of the items would be improved by deleting or revising response options.

In set 1 (the only set applied in both Uganda and Norway), five items displayed uniform DIF and one item displayed non-uniform DIF associated with setting (i.e., Norway and Uganda). Across the four sets including 88 items, six items displayed uniform DIF associated with age, and 7 items indicated DIF associated with reading ability.

Multidimensionality and response dependency

We did not observe any statistically significant residual correlation between pairs of items, and the paired *t*-test procedure indicated that the sets were sufficiently unidimensional (Table 2).

Discussion

We have developed the Claim Evaluation Tools database using qualitative and quantitative feedback from methodologists and end-users in six countries (23). This study reports the findings of the first psychometric testing of multiple-choice items from the database, using Rasch analysis conducted in two settings, Uganda and Norway.

Most of the 88 items (addressing 22 Key Concepts) conformed well to the Rasch model. However some items displayed DIF and required revisions. Overall, we found that the four item sets created from the Claim Evaluation Tools database had acceptable reliability. We did not identify significant response dependence between any pairs of items and, the magnitude of multidimensionality in the data was acceptable.

Based on findings from previously reported descriptive and qualitative methods, experts and end-users suggested that the items were potentially difficult for members in our target group (23). The Rasch analyses confirmed this. Furthermore, using the reading items we developed, the respondents reading skills were found to be low. It should be noted that these items have not previously been tested, and served as a pragmatic indicator of the respondents' ability to identify and apply the correct text in response to questions relating to a scenario similar to what we use in the multiple-choice items. We also explored DIF by reading ability. Only seven out of 88 items displayed evidence of DIF, suggesting that most items work in the same way independent of people's reading ability as measured in this study.

This suggests that efforts should be made to simplify the text in the scenarios and editing the response options to improve readability and improve validity. Reducing the number of response options in the items could also contribute to making the items less difficult.

Based on the findings from our analyses, we decided to remove items with non-uniform DIF from the Claim Evaluation Database. We also decided to revise items with poor model fit and reduce response options that did not work as expected. Revised items will be retested in the Ugandan context, where they will be used as an outcome measure.

A limitation of this study is that we tested the items in only two settings, Uganda and Norway, and that the fit to the Rasch model in other settings is unknown. Further testing of items from the Claim Evaluation Tools database using Rasch Analysis in other countries and languages is needed. We also did not include the respondents' gender in the analysis, which could introduce further DIF. This will be explored in further testing.

There has been an encouraging interest in the Claim Evaluation Tools database in settings other than the countries included in the IHC project, and researchers in Norway, Mexico, Germany and China are currently translating and testing the multiple-choice items in their settings. In addition, the items addressing the Key Concepts we judged to be more advanced, and which were not tested as part of this study, are currently being tested online through www.testingtreatments.org, targeting people with relevant training, such as health researchers or teachers of evidence-based medicine. We are also developing items to assess intended behaviours and attitudes towards assessing treatment claims. The Claim Evaluation Tools database, which includes all of these questions, as well as findings from evaluations such as this one, is freely available for non-commercial use on request through the Testing Treatments interactive website (www.testingtreatments.org).

When used for evaluating peoples' ability to assess treatment claims, an item set generated from the Claim Evaluation Tools database can be scored by calculating the number or percentage of correct responses. However, such scores can be difficult to interpret, especially when comparing the average score of two groups (e.g. in a randomised trial). An absolute (criterion referenced) standard for a passing score (i.e. a cut-off for passing) or for mastery of the Key Concepts that are tested. Setting a cut-off requires judgement, and there are several ways of doing this (43-45). For the items that will be used in the trials of the IHC primary school resources and podcast, we have established criteria referenced standards using a combination of Nedelsky's and Angoff's methods (43-46).

Conclusion

We found that most items that we tested had satisfactory fit to the Rasch model. Taken together with our previously reported findings, the findings of this study suggest that the items have face and construct validity in the settings in which they have been tested. Following revisions of some items, informed by the findings from this study, most of the items that we tested are suitable for use in an outcome measure that evaluates people's ability to apply the key concepts they need to know to be able to assess treatment claims.

Authors' contributions

AA, ØG, AO wrote the protocol and the IHC group provided comments to this protocol. The Claim Evaluation Tools database was developed by the IHC group. AA coordinated all of the development and evaluation process with support from AO. DS, AN and KO performed the data collection and data entry from the field-testing. ØG and AA prepared the data files for the analysis, and conducted the Rasch analyses. AA authored this manuscript with significant input from the IHC group.

Acknowledgements

We are deeply grateful to all of the enthusiastic children, parents and teachers that contributed to this project. We would also like to thank the IHC advisory panel, and the other researchers and methodologists that provided their advice in the development process.

Funding and competing interests

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

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.

Data sharing statement

All data are published as part of this study; additional information is available upon request and on our website informedhealthchoices.org. All items in the Claim Evaluation Tools Database are available upon request for non-commercial use.

References

1. Lewis M, Orrock P, Myers S. Uncritical reverence in CM reporting: Assessing the scientific quality of Australian news media reports. Health Sociology Review. 2010;19(1):57-72.
2. Glenton C, Paulsen E, Oxman A. Portals to Wonderland? Health portals lead confusing information about the effects of health care. BMC Medical Informatics and Decision Making. 2005;5:7:8.
3. Moynihan R, Bero L, Ross-Degnan D, Henry D, Lee K, Watkins J, et al. Coverage by the news media of the benefits and risks of medications. The New England Journal of Medicine. 2000;342(22):1645-50.
4. Wolfe R, Sharp L, Lipsky M. Content and design attributes of antivaccination web sites. Journal of American Medical Association. 2002;287(24):3245-48.
5. Woloshin S, Schwartz L, Byram S, Sox H, Fischhoff B, Welch H. Women's understanding of the mammography screening debate. Archives of Internal Medicine 2000;160:1434-40.
6. Fox S, Duggan M. Health Online 20132013 09.04.2013. Available from: http://www.pewinternet.org/Reports/2013/Health-online.aspx.
7. Robinson E, Kerr C, Stevens A, Lilford R, Braunholtz D, Edwards S, et al. Lay public's understanding of equipoise and randomisation in randomised controlled trials. Research Support, Non-U.S. Gov't. NHS R&D HTA Programme, 2005 Mar. Report No.: 1366-5278 (Linking) Contract No.: 8.
8. Sillence E, Briggs P, Harris PR, Fishwick L. How do patients evaluate and make use of online health information? Social Science & Medicine. 2007;64(9):1853-62.
9. Horsley T, Hyde C, Santesso N, Parkes J, Milne R, Stewart R. Teaching critical appraisal skills in healthcare settings. Cochrane Database of Systematic Reviews. 2011(11).
10. Evans I, Thornton H, Chalmers I, P. G. Testing Treatments: better research for better healthcare. Second edition. London: Pinter & Martin Ltd2011. Available from: Available online at www.testingtreatments.org/new-edition/.
11. Chalmers I., Glasziou P., Badenoch D., Atkinson P., Austvoll-Dahlgren A., Oxman A. Evidence Live 2016: Promoting informed healthcare choices by helping people assess treatment claims. BMJ; 26.06.2016.

12. Taking shared decision making more seriously. Lancet. 2011;377(9768):784.

13. Stacey D, Légaré F, Col NF, Bennett CL, Barry MJ, Eden KB, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database of Systematic Reviews 2014, Issue 1 Art No: CD001431 DOI: 101002/14651858CD001431pub4.

14. Berkman N, Sheridan S, Donahue K, Halpern D, Crotty K. Low Health Literacy and Health Outcomes: An Updated Systematic Review. Annals of Internal Medicine. 2011;155(2):97-U89.

15. 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. Systematic Reviews 2016;5:215.

16. Abrami PC, Bernard RM, Borokhovski E, Waddington DI, Wade CA, T. P. Strategies for Teaching Students to Think Critically: A Meta-Analysis. Review of Educational Research June 2015, Vol 85, No 2, pp 275– 314.

17. Sorensen K, Van den Broucke S, Fullam J, Doyle G, Pelikan J, Slonska Z, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health. 2012;12:80.

18. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312(7023):71-2.

19. Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol. 2011;64(4):380-82.

20. 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. Submitted manuscript. 2016.

21. 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 claims about the benefits and harms of treatments? Protocol for a randomized trial

Submitted manuscript. 2016.

22. 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. Journal of Evidence-Based Medicine. 2015;8(3):112-25.

23. 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". Accepted manuscript BMJ open. 2016.

24. Case S, Swanson D. Constructing Written Test Questions For the Basic and Clinical Sciences (Third edition). Philadelphia, USA: 2002.

25. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol. 2007;6(12):1094-105.

26. Nsangi A, Semakula D, Oxman AD, Sewankambo NK. Teaching children in low-income countries to assess claims about treatment effects: prioritization of key concepts. Journal of Evidence-Based Medicine. 2015;8(4):173-80.

27. Semakula D, Nsangi A, Oxman AD, Sewankambo NK. Priority setting for resources to improve the understanding of information about claims of treatment effects in the mass media. Journal of Evidence-Based Medicine. 2015;8(2):84-90.

28. Linacre J. Sample size and item calibration stability. Rasch Measurement Transactions. 1994;7:328.

29. Lexile-to-Grade Correspondence. MetaMetrics. 2017. https://www.lexile.com/about-lexile/grade-equivalent/grade-equivalent-chart/

30. Leonard M. Rasch Promises: a Layman's Guide to the Rasch Method of Item Analysis. Educational Research. 1980;22(3):188-92.

31. 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):1358-62.

32. Guttersrud O, Dalane JO, Pettersen S. Improving measurement in nutrition literacy research using Rasch modelling: examining construct validity of stage-specific 'critical nutrition literacy' scales. Public Health Nutr. 2014;17(4):877-83.

33. Conaghan PG, Emerton M, Tennant A. Internal construct validity of the Oxford Knee Scale: evidence from Rasch measurement. Arthritis Rheum. 2007;57(8):1363-7.

34. Rasch analysis. http://www.rasch-analysis.com/. Accessed 2016.

35. 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.

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

37. Displaying the RUMM2030 Analysis. Rasch Unidimensional Measurement Model. 2015.

38. Brodersen J, Meads D, Kreiner S, Thorsen H, Doward L, McKenna S. Methodological aspects of differential item functioning in the Rasch model. J Med Econ. 2007;10:309 – 24.

39. Marais I, Andrich D. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. J Appl Meas. 2008;9(3):200-15.

40. RUMM. Extending the RUMM2030 Analysis. 7. ed: RUMM Laboratory Pty Ltd. 2009.

41. 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 Journal of Statistics. 2014

42. Andrich D., Humphry SM., Marais I. Quantifying local, response dependence between two polytomous items using the. Applied Psychological Measurement 36(4), 309–324. 2012.

43. Livingston SA ZM. Passing scores; A manual for setting standards of performance on educational and occupational tests. Educational Testing Service. 1982.

44. Nedelsky L. Absolute grading standards for objective tests. Education and Psycholgical Measurement Journal. 1954;14(1):3-19.

45. Angoff WH. 4. Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL (ed.). Educational Measurement Washington DC. 1971:514-5.

46. Davies A., Gerrity M., Nordheim L., Okebukola P., Opiyo N., Sharples J., et al. Measuring ability to assess claims about treatment effects: establishment of a standard for passing and mastery. IHC Working Paper 2017; ISBN 978-82-8082-802-6.

Figure legends

Figure 1. Example of a multiple choice-item taken from the Claim Evaluation Tools database

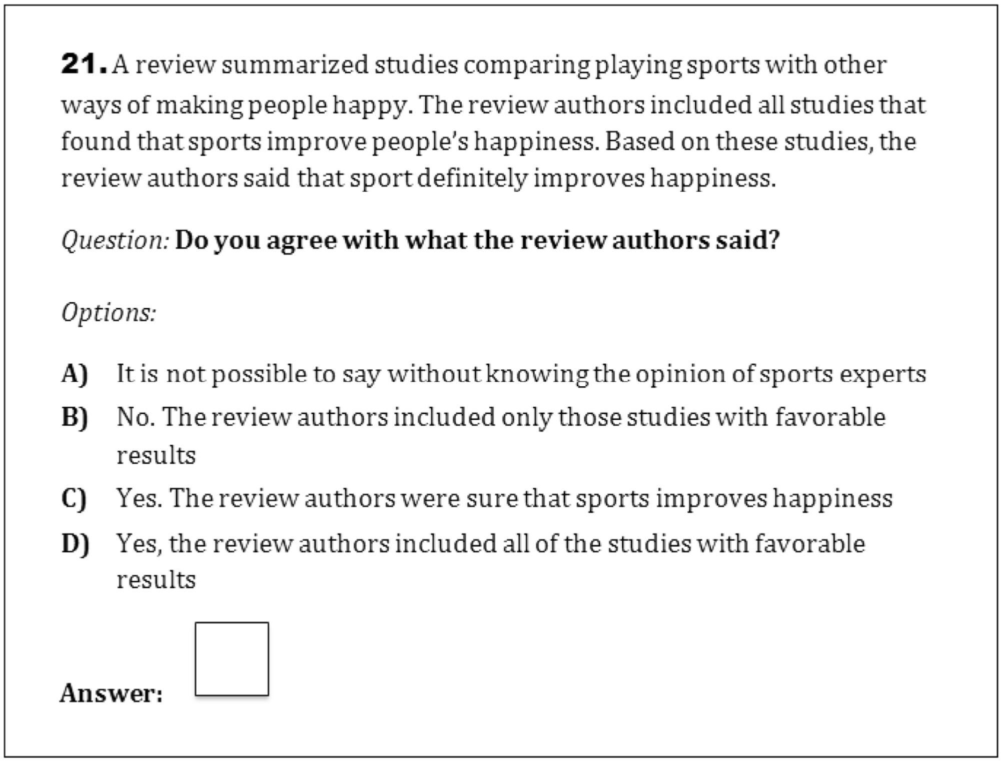
Figure 2. The Item Characteristic Curve

Figure 3. Item Person map for item set 1

Figure 4. Item Person map for item set 2.

Figure 5. Item Person map for item set 3

Figure 6. Item Person map for item set 4



21. A review summarized studies comparing playing sports with other ways of making people happy. The review authors included all studies that found that sports improve people's happiness. Based on these studies, the review authors said that sport definitely improves happiness.

Question: **Do you agree with what the review authors said?**

Options:

A) It is not possible to say without knowing the opinion of sports experts

B) No. The review authors included only those studies with favorable results

C) Yes. The review authors were sure that sports improves happiness

D) Yes, the review authors included all of the studies with favorable results

Answer: □

Figure 1. Example of a multiple choice-item taken from the Claim Evaluation Tools database

162x123mm (300 x 300 DPI)

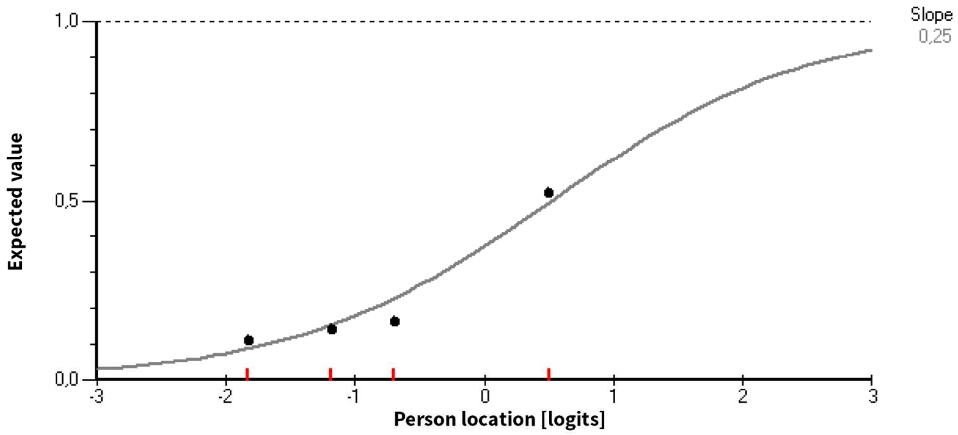


Figure 2. The Item Characteristic Curve

163x82mm (300 x 300 DPI)

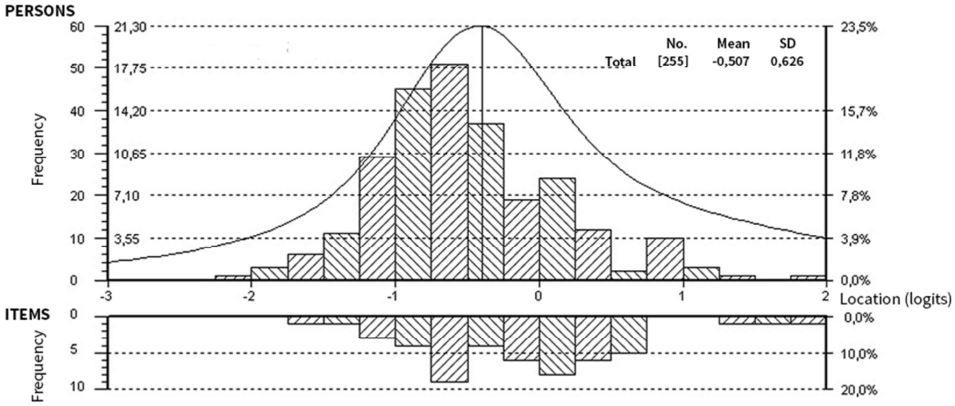


Figure 3. Item Person map for item set 1

82x37mm (300 x 300 DPI)

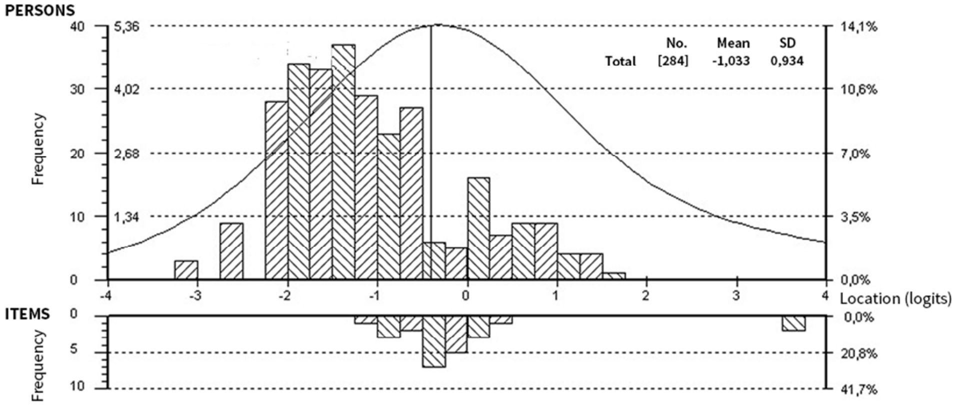


Figure 4. Item Person map for item set 2.

82x37mm (300 x 300 DPI)

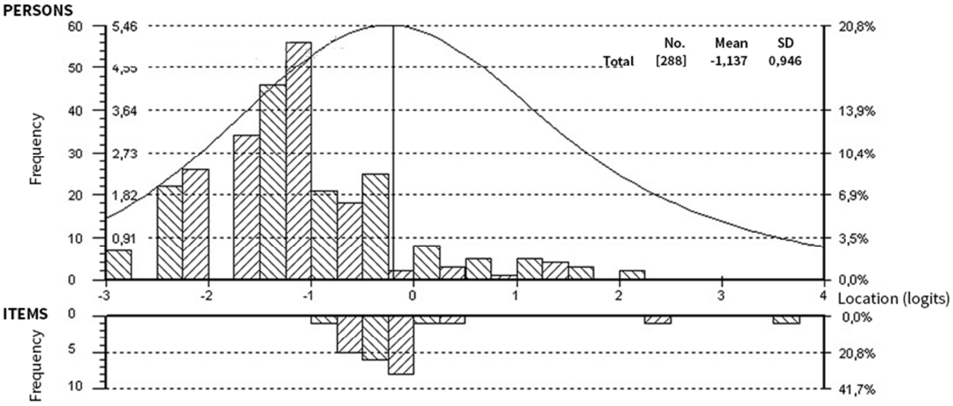


Figure 5. Item Person map for item set 3

82x37mm (300 x 300 DPI)

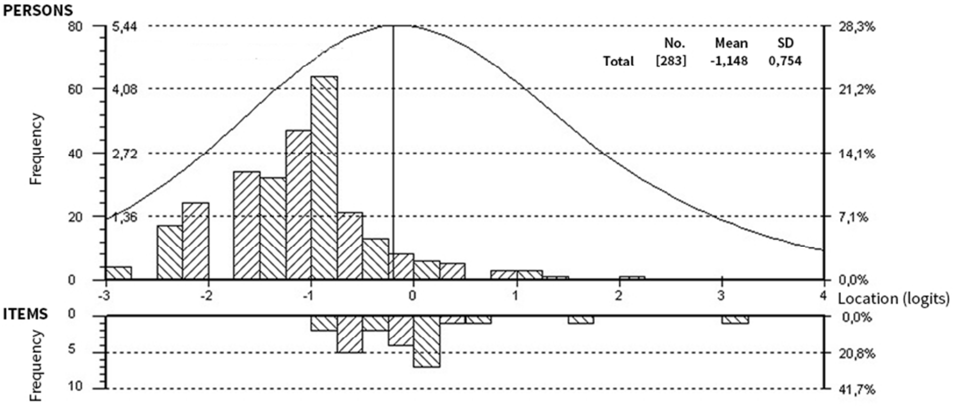


Figure 6. Item Person map for item set 4

82x37mm (300 x 300 DPI)

1. A doctor did a study to find out if drinking tea keeps people from getting sick. The doctor tossed a coin to decide who should get the tea and who should not. People who got tea went to the doctor's office every day to drink their tea. At the end of the study, people who got the tea were less likely to be sick than those who got no tea.

Based on the text above, please answer the following questions:

1.1 Who went to the doctor's office every day?

Options:

A) People who did not get tea

B) People who got tea

C) Everyone

D) People who got sick

Answer: □

1.2. How did the doctor decide who should get tea?

Options:

A) By tossing a coin

B) By asking people what they would like

C) They gave tea to those who were more likely to be sick

D) They asked people who came to their office

Answer: □

1.3. What was the treatment?

Options:

A) Tea

B) Sleep

C) The study

D) The doctors

Answer: □

1.4. What was the result of the study?

Options:

A) Drinking tea can help people from getting sick

B) Doctors should toss coins when doing studies

C) People should go to the doctor if they are sick

D) Not drinking tea can help people from getting sick

Answer: □