PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Measuring ability to assess claims about treatment effects: A latent trait analysis of items from the “Claim Evaluation Tools” database using Rasch modelling
AUTHORS	Austvoll-Dahlgren, Astrid; Guttersrud, Øystein; Nsangi, Allen; Semakula, Daniel; Oxman, Andrew

VERSION 1 - REVIEW

REVIEWER	Dr Susan Darzins Australian Catholic University, Australia
REVIEW RETURNED	17-Aug-2016

GENERAL COMMENTS	Thank you for the opportunity to review this manuscript. The authors have chosen a rigorous approach to internal construct validation of a scale to be used in two randomised controlled trials. Concepts related to Rasch analysis are difficult to convey and the authors have done this clearly. Apart from a couple of typos I make suggestions for more complete reporting of the Rasch analysis procedures, and results, to aid transparency of the process and the evidence that was generated by the research: 1. Page 4 under the heading 'Strengths and limitations of this study': - in the fifth point, I suggest change of wording to " The items tested in this study were tested" (rather than 'was' tested). 2. Page 6, third line from the bottom I suggest change of wording to "The Claim Evaluation Tools were developed in English, but are currently being translated" (i.e. use 'are' rather than 'is') 3. Page 7, first and second lines from the top I suggest change of wording to improve readability to "the comparison of two people is independent of which items...." (i.e. replace 'are' with 'is') 4. Page 12 under heading 'The components of Rasch Analysis': It would be useful for increased transparency and rigour in the reporting of the methods if the authors reported the criteria they set as acceptable/not acceptable in the data, for all of the Rasch procedures, which would have informed their decisions about the scales. For example, when evaluating local independence, what magnitude of correlation coefficient was accepted/was considered to violate local independence? Another example: what cut-off value for the PCA/t-test procedure was used, and was the value's 95%CI used? This information could be presented quite nicely in a Table format. 5. The authors could check that all data analysis procedures are mentioned in the methods section. For example, there was no

	mention that Cronbach's alpha would be used as a test of reliability, but was then reported in the results section. Could the authors provide a rationale for use of Cronbach's alpha as a test of reliability during Rasch analysis, rather than the Person Separation Index, which available in RUM2030. 6. Page 13 under the heading 'Results': Some results were presented in a relatively general way, for example "Most of the items conformed well to the Rasch model and only a few items showed evidence of DIF. The readers require knowledge about the criteria the authors used for making these decisions, and they also need to know what the results in the data were, for each of the criteria, to be able to accept the evidence as reported. It was not clear why the authors did not report statistics such as the overall model fit (Chi-square score, df probability value), overall item fit residual statistic and its SD, overall person fit residual statistic and its SD, number of misfitting items, number of misfitting persons, Person Separation Index scores, whether the DIF observed was uniform or non-uniform, the number of item pairs with local response dependency (and the values) and the PCA/t-test percentage of significant t-tests. These could be presented in Table 3. 7. In Table 3, could the authors note what NR means 8. Page 13, under the heading 'Targeting and reliability', could the authors clarify which logits are being reported and what type of spread was expected in the variable 'ability'. 9. Page 13 under the heading 'Possible dimension and response violation of local independence', could the authors be more specific (as already mentioned) in the reporting as to how the data in the four sets were deemed to measure a sufficiently unidimensional latent trait. 10. It would be useful to see item maps, or item-person maps for the scales in each of the groups. 11. Could the authors carefully proof-read the manuscript again for grammar related to plural/singular terms and for insertion of commas to improve readability. 12. Ethics approval. Assurance that ethics approval was 'received' rather than 'sought' would strengthen the statement. Please check if it is a requirement to provide ethics approval numbers for the data collection sites. 13. If there is further clarification of the analysis methods and results then it is possible that the discussion and conclusions may be justified by the results.

REVIEWER	Levente Kriston University Medical Center Hamburg-Eppendorf, Germany
REVIEW RETURNED	28-Nov-2016

GENERAL COMMENTS	The authors report on the psychometric testing of the “Claim Evaluation Tools“, which was developed to measure ability to assess

	claims about treatment effects. Using Rasch modelling in samples from Uganda and Norway, they conclude that the measure has sufficient reliability but should be tested in further settings. Although it deals with an important topic, I had some difficulties to understand the report. Major comments: 1. I missed a clear definition of the central terms in the manuscript, such as “claim” (e.g., p. 5, line 7), “Key concepts” (e.g., p. 5, line 44), “relevance for specific populations” (e.g., p. 6,. line 22) and “critically assess” (p. 7, line 54). In general, the construct that is measured should be defined in more detail, including interpretation of high and low scores, probably using terminology from ability/competence/performance measurement. For example, if the authors assume that there are correct answers to the questions and that they concern understanding information, the construct of interest might be related to health literacy, knowledge, or even general intelligence. Therefore, a clear distinction form related constructs should be made (validity). 2. A major barrier to understanding the report was the lack of information provided on previous work. As the most important references (references 16, 17, 19 and 20) are all “submitted” (and probably should be removed from the reference list), it is essentially impossible to get information on the history of the measure (including existing findings). The missing information should be added to the present manuscript in order to make it more comprehensible. 3. More comprehensive information on the content of the measure and the context of its assessment should be provided. For example, although not an explicit reporting guideline, checking the COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN; http://www.cosmin.nl/) may give an impression of the information needed for a thorough assessment of measures. 4. The basics of Rasch analysis are presented very extensively (p. 7 to p. 10). The methods could be summarized in a more pregnant way. Instead, more attention should be given to the Results and their interpretation. 5. P. 11, line 35 ff. It is incomprehensibly presented what the 32 key concepts are, how 22 of them were selected, and how and why four subsets of them were built (e.g. were the items randomly allocated to the subsets?). More details should be provided, including how many items were included in the subsets (Table 1). 6. More information on the sample (including descriptive statistics) should be given. 7. Far more numerical and graphical results of the analyses should be presented, also on the item-level, including complete information on thresholds, fit statistics, differential item functioning, standard errors, information function, item-person map, item characteristic (category probability) curves, factor loadings, information functions etc. Otherwise, it is not possible to assess whether the conclusions are supported by the data.

	8. Differential item functioning by literacy should be checked. In addition, association of the final score with literacy should be investigated and reported, in order to be able to assess whether the measure's target is sufficiently distinct from literacy (see also comment 1). 9. All items of the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) statement should be addressed adequately (http://strobe-statement.org). Minor comments: 10. Throughout the manuscript, it remained unclear to me why the authors speak about tools and instruments in plural. As for my understanding, it is a single tool with multiple items and subsets that was tested. This should be more clearly described. 11. p. 3 line 12: It was unclear, what is meant by "flexible" items? Do the authors mean something like a set, from which subsets can be flexibly chosen? 12. p. 3 line 16: At least one subheading seems to be missing, as after the first sentence the text is not on "Participants". 13. p. 13, line 40. It is unclear what is meant by the statement that the spread in ability was "as expected". Specifications should be provided. 14. p. 14, line 51. Differential item functioning by gender could be easily included in the manuscript. 15. The Tables and Figures frequently lack a sufficiently comprehensible legend and/or a description of what they present. 16. p. 17, line 21. The specific institutions/review boards approving the study protocol should be named. 17. The language should be improved.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

1. Page 4 under the heading 'Strengths and limitations of this study':
- in the fifth point, I suggest change of wording to " The items tested in this study were tested" (rather than 'was' tested).

Response: We agree that this sentence was a bit awkward. We have merged the two last sentences to improve clarity.

2. Page 6, third line from the bottom I suggest change of wording to "The Claim Evaluation Tools were developed in English, but are currently being translated" (i.e. use 'are' rather than 'is')

Response: This has been corrected.

3. Page 7, first and second lines from the top I suggest change of wording to improve readability to "the comparison of two people is independent of which items...." (i.e. replace 'are' with 'is')

Response: This has been corrected.

4. Page 12 under heading 'The components of Rasch Analysis':
It would be useful for increased transparency and rigour in the reporting of the methods if the authors reported the criteria they set as acceptable/not acceptable in the data, for all of the Rasch procedures, which would have informed their decisions about the scales. For example, when evaluating local independence, what magnitude of correlation coefficient was accepted/was considered to violate local independence? Another example: what cut-off value for the PCA/t-test procedure was used, and was the value's 95%CI used? This information could be presented quite nicely in a Table format.

Response: This has been corrected, and the methods section has undergone major revision to make it more precise and to improve transparency. The methods section has been restructured to follow the fundamental steps of Rasch analysis, and all cut-off values are now presented.

5. The authors could check that all data analysis procedures are mentioned in the methods section. For example, there was no mention that Cronbach's alpha would be used as a test of reliability, but was then reported in the results section. Could the authors provide a rationale for use of Cronbach's alpha as a test of reliability during Rasch analysis, rather than the Person Separation Index, which available in RUM2030.

Response: We now report both, and the two reliability measures are now described in the methods section.

6. Page 13 under the heading 'Results':
Some results were presented in a relatively general way, for example "Most of the items conformed well to the Rasch model and only a few items showed evidence of DIF. The readers require knowledge about the criteria the authors used for making these decisions, and they also need to know what the results in the data were, for each of the criteria, to be able to accept the evidence as reported. It was not clear why the authors did not report statistics such as the overall model fit (Chi-square score, df probability value), overall item fit residual statistic and its SD, overall person fit residual statistic and its SD, number of misfitting items, number of misfitting persons, Person Separation Index scores, whether the DIF observed was uniform or non-uniform, the number of item pairs with local response dependency (and the values) and the PCA/t-test percentage of significant t-tests. These could be presented in Table 3.

Response: This has been corrected, and results section is now structured similarly as the new methods section representing the fundamental steps of Rasch analysis. All overall and individual Fit statistics are now presented in text or in Table 2. We also added Item Person maps for all sets.

7. In Table 3, could the authors note what NR means

Response: The Norwegian sample was small and we were only able to test one of the four sets of items in this setting. The description of DIF has been added to the text and the table has been deleted.

8. Page 13, under the heading 'Targeting and reliability', could the authors clarify which logits are being reported and what type of spread was expected in the variable 'ability'.

Response: This has been corrected. See also comment 4.

9. Page 13 under the heading 'Possible dimension and response violation of local independence',

could the authors be more specific (as already mentioned) in the reporting as to how the data in the four sets were deemed to measure a sufficiently unidimensional latent trait.

Response: his has been corrected. See also comment 4.

10. It would be useful to see item maps, or item-person maps for the scales in each of the groups.

Author's comment: We have added item maps to the manuscript as Figures 3 to 6.

11. Could the authors carefully proof-read the manuscript again for grammar related to plural/singular terms and for insertion of commas to improve readability.

Response: We have proof-read the manuscript again.

12. Ethics approval. Assurance that ethics approval was 'received' rather than 'sought' would strengthen the statement. Please check if it is a requirement to provide ethics approval numbers for the data collection sites.

Response: We have clarified this in the manuscript.

Reviewer: 2

1. I missed a clear definition of the central terms in the manuscript, such as "claim" (e.g., p. 5, line 7), "Key concepts" (e.g., p. 5, line 44), "relevance for specific populations" (e.g., p. 6,. line 22) and "critically assess" (p. 7, line 54). In general, the construct that is measured should be defined in more detail, including interpretation of high and low scores, probably using terminology from ability/competence/performance measurement. For example, if the authors assume that there are correct answers to the questions and that they concern understanding information, the construct of interest might be related to health literacy, knowledge, or even general intelligence. Therefore, a clear distinction form related constructs should be made (validity).

Response: This paper was submitted alongside another paper (now accepted by BMJ open) describing the development of the items, however we agree that some information describing the concepts underlying the items and initial steps of the developments should also be mentioned briefly in this paper. This has been added to the manuscript.

2. A major barrier to understanding the report was the lack of information provided on previous work. As the most important references (references 16, 17, 19 and 20) are all "submitted" (and probably should be removed from the reference list), it is essentially impossible to get information on the history of the measure (including existing findings). The missing information should be added to the present manuscript in order to make it more comprehensible.

Response: This has been added. See also response to reviewer 2's second comment.

3. More comprehensive information on the content of the measure and the context of its assessment should be provided. For example, although not an explicit reporting guideline, checking the COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN; http://www.cosmin.nl/) may give an impression of the information needed for a thorough assessment of measures.

4. Response: We believe our methods and reporting of results now adhere to the COSMIN criteria. We also reference COSMIN as part of the background section.

4. The basics of Rasch analysis are presented very extensively (p. 7 to p. 10). The methods could be summarized in a more pregnant way. Instead, more attention should be given to the Results and their interpretation.

Response: We have simplified this description greatly to improve clarity. See also our response to reviewer 1's 4th comment.

5. P. 11, line 35 ff. It is incomprehensibly presented what the 32 key concepts are, how 22 of them were selected, and how and why four subsets of them were built (e.g. were the items randomly allocated to the subsets?). More details should be provided, including how many items were included in the subsets (Table 1).

Response: A description of how the items were selected has been added to the methods section.

5. More information on the sample (including descriptive statistics) should be given.

Response: This has been added to the methods section.

7. Far more numerical and graphical results of the analyses should be presented, also on the item-level, including complete information on thresholds, fit statistics, differential item functioning, standard errors, information function, item-person map, item characteristic (category probability) curves, factor loadings, information functions etc. Otherwise, it is not possible to assess whether the conclusions are supported by the data.

Response: This has been corrected; see also our response to reviewer 1's 4th comment. However, category probability curves were not included, as this is not relevant to dichotomous items.

8. Differential item functioning by literacy should be checked. In addition, association of the final score with literacy should be investigated and reported, in order to be able to assess whether the measure's target is sufficiently distinct from literacy (see also comment 1).

Response: Literacy was not investigated as a factor causing DIF, this was incorrectly described in the methods section and has been corrected. However, we include this as background information of the respondents' text recognition and understanding, which we used as an indication of reading ability.

9. All items of the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) statement should be addressed adequately (http://strobe-statement.org).

Response: The manuscript now captures criteria: 1, 2, 3, 4, 5, 6, 9, 10 in the STROBE checklist for cross-sectional studies. Since the purpose of this study was to perform psychometric testing of the items - not to map or generalize the results of the questionnaires (the respondents' ability to assess treatment claims) - the other STROBE criteria are not relevant for this study.

10. Throughout the manuscript, it remained unclear to me why the authors speak about tools and instruments in plural. As for my understanding, it is a single tool with multiple items and subsets that was tested. This should be more clearly described.

Response: The Claim Evaluation Tools database includes a battery of multiple-choice questions that

can be used in a variety of evaluation tools. We have clarified this in the manuscript and only use the term “tools” when referring to the database and we only use the term “instrument” (singular) once, when reporting that we were unable to find a suitable instrument.

11. p. 3 line 12: It was unclear, what is meant by “flexible” items? Do the authors mean something like a set, from which subsets can be flexibly chosen?

Response: Our vision for the Claim Evaluation Tools was that they should not be a standard, fixed questionnaire, but rather a flexible tool-set including a battery of items, from which a set of relevant items can be selected forspecific populations and purposes. For example, a teacher developing a series of lectures targeting five of the concepts in the Key Concept list, could design her own evaluation instrument (tool) to test her students by picking items from the database that specifically address those Key Concepts. We have added a description to the text to clarify this.

12. p. 3 line 16: At least one subheading seems to be missing, as after the first sentence the text is not on “Participants”.

Response: This has been corrected.

13. p. 13, line 40. It is unclear what is meant by the statement that the spread in ability was “as expected”. Specifications should be provided.

Response: This has been corrected; see also our response to reviewer 1’s 4th comment.

14. p. 14, line 51. Differential item functioning by gender could be easily included in the manuscript.

Response: This information was not collected as part of this study - but has been included in further testing. These results have not been published yet, but based on testing in two other contexts we have not found evidence of DIF by gender.

15. The Tables and Figures frequently lack a sufficiently comprehensible legend and/or a description of what they present.

Response: This has been corrected.

16. p. 17, line 21. The specific institutions/review boards approving the study protocol should be named.

Response: This has been specified.

17. The language should be improved.

Response: We have proof-read the manuscript again.

VERSION 2 – REVIEW

REVIEWER	Levente Kriston University Medical Center Hamburg-Eppendorf, Germany
REVIEW RETURNED	17-Feb-2017

GENERAL COMMENTS	The authors addressed my comments appropriately, and I think they did a great job in revising the manuscript in general. I have only one concern left: Some evidence should be provided that the ability to assess claims about treatment effect can be sufficiently separated from general literacy. In my opinion, the level of association between the two measures should be reported (quantitatively, e.g., as a correlation coefficient), or the limitation that construct validity is questionable due to unclear distinction form general literacy should be explicitly included in the Discussion section.

VERSION 2 – AUTHOR RESPONSE

Thank you for accepting our revisions. We have done our best to accommodate the last request from the peer-reviewer.

Literacy can be assumed to affect peoples' responses to all instruments administrated as written questionnaires, including tests evaluating knowledge and instruments assessing attitudes, satisfaction or self-reported behavior. Although this is seldom explored or recognized. Literacy may thus introduce measurement error (such as differential item functioning) or act as a covariate in observational or experimental studies.

We knew literacy would be a potential problem in our target audience, and wanted to test this using four items we developed for this purpose. However, it should be noted that these items have not been previously evaluated. Furthermore, this measurement served only 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.

Furthermore, the purpose of this study was not to measure people's ability to apply the key concepts in a representative sample, but to evaluate the validity and reliability of the four sets using Rasch analysis. Thus exploring literacy as a covariate would not fit within the scope or objectives of this study. However, reading ability could potentially introduce differential item function. Therefore, in response to the reviewer's request, we have analysed the invariance to the scales on item level. Although there is no "gold standard" for how to best analyse DIF using continuous variables such as our reading ability items, we have now performed an analysis where we recode the scores of these items into a categorical variable. The results of this analysis indicates that DIF by literacy was not an important problem, and that only 7 out of 88 items showed signs of DIF. All of these were uniform. The results and implications of this analysis is now also described in the discussion.

Finally, we would like to take the opportunity to thank the peer-reviewers again for their very useful feedback in this process. We believe it has improved our manuscript.