PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Measuring ability to assess claims about treatment effects: The development of the “Claim Evaluation Tools”
AUTHORS	Austvoll-Dahlgren, Astrid; Semakula, Daniel; Nsangi, Allen; Oxman, Andrew; Chalmers, Iain; Rosenbaum, Sarah; Guttersrud, Øystein

VERSION 1 - REVIEW

REVIEWER	Susan Darzins Australian Catholic University
REVIEW RETURNED	24-Jul-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	Shaun Treweek University of Aberdeen, UK I have worked with some of the authors in the past.
REVIEW RETURNED	25-Jul-2016

GENERAL COMMENTS

Introduction
This is a well written article describing a tool comprising a set of multiple-choice questions to measure people's ability to understand and apply some key concepts needed to assess claims about treatment effects. As the authors state, this is likely to be the first such tool.

I have a few minor comments, which are given below.

General
1. I think there is a difference between 'key concepts' and 'Key Concepts'. The former is very general, the latter suggests a specific set of concepts. The authors clearly do have a list that they could refer to as Key Concepts (table 1) but I think they should start out in their article by referring to key concepts and then moved to saying that they have a set list of concepts in mind and that these will now be called Key Concepts. The authors might think of a better way of doing it this but my key point is that the first few references two key concepts should imply something general before the authors move to something more specific.

Strengths and limitations
1. I may have misunderstood but I think I'm right in saying that the paper presents validation and pilot testing of 22 of 32 items presented in table 1. I Think the authors need to say somewhere, probably in the Discussion, but also in a bullet point here in Strengths and limitations what this means for the remaining 10 items. Should I feel as comfortable using them as the other 22?
2. I would use 'key concepts' here and not the capitalised version. The first bullet item also needs a 'that' before '.. people need to know..'.
'relevant' in the third bullet point should be 'relevance'.

Results
1. Page 12, line 18 'party' under face validity should be 'partly'.
2. Page 13, under 'Preference of format and missing responses' I think it would be good to have some examples of comments coming from the interviews with end users that support the summary given. This can either be a table or box with example comments, or perhaps a list of all of them in a supplementary document.
3. Page 13: I would re-order the last sentence of the second paragraph to something like 'Figure 2 shows the design changes used to avoid these problems'. Also, are the changes shown in figure 2 all the changes that were made to handle missing data, or were there others?
Page 14, first paragraph. 'figure 3' should be 'figure 4'.

Discussion
1. The background section says that the authors set out to develop the Claim Evaluation Tools to serve as the primary outcome measures to be used in future trials. The discussion section does not make explicit reference to how suitable or otherwise the authors now think the Tools are for this purpose. It would be useful for the authors to add a sentence or two that makes explicit reference to this aim.
2. As mentioned earlier, I think the authors need to say something about the 10 items not tested (or make the text clearer so that my misunderstanding is not repeated by other readers).

	3. Page 14, line 26: there is an ‘a’ missing in ‘.. We believe that they will be A useful tool..’ 4. Page 15: ‘.. measuring literacy skills..’ - do the authors have any thoughts on what level of literacy is required to successfully use the Tools in future evaluations? At present the authors say it would be important to measure literacy skills but do not say what level of minimum literacy is required to successfully use the Tools. Table 1 1. I think the item headings (e.g. 1. recognising the need for fair comparisons of treatments’) should be bolded, in italics or some similar formatting change so that it is clear that these are headings not items themselves. 2. I think the caption should be explicit fact 22/32 were tested here (or correct my misunderstanding).

VERSION 1 – AUTHOR RESPONSE

Reviewer: Shaun Treweek

General
1. I think there is a difference between ‘key concepts’ and ‘Key Concepts’. The former is very general, the latter suggests a specific set of concepts. The authors clearly do have a list that they could refer to as Key Concepts (table 1) but I think they should start out in their article by referring to key concepts and then moved to saying that they have a set list of concepts in mind and that these will now be called Key Concepts. The authors might think of a better way of doing it this but my key point is that the first few references two key concepts should imply something general before the authors move to something more specific.
Author’s feedback: We have revised the introduction to make this distinction more clear.

Strengths and limitations
1. I may have misunderstood but I think I’m right in saying that the paper presents validation and pilot testing of 22 of 32 items presented in table 1. I Think the authors need to say somewhere, probably in the Discussion, but also in a bullet point here in Strengths and limitations what this means for the remaining 10 items. Should I feel as comfortable using them as the other 22?
Author’s feedback: This paper describes the development and validation of items addressing all of the 32 Key Concepts including four phases. In the last phase, we also did some pilot testing for which items referring to 22 of the 32 Key Concepts were included. The purpose of these pilots were to do practical administrative tests to explore understanding of formats and timing of a “sample test”, but also to get some kind of indication of the sample size needed for the IHC trials. Which Key Concepts were targeted in this test were judged to be of little importance as the items addressing the different key Concepts use the same formats and are equal in length and language. It is however important to note that this paper describes the development, judgements of face validity (content validity) and judgements about relevance. To what extent the items are reliable cannot be judged based on the data reported in this study, but is addressed in a separate paper. To make this more explicit, we have revised the section describing the next phase of psychometric testing in the discussion.

2. I would use ‘key concepts’ here and not the capitalised version. The first bullet item also needs a ‘that’ before ‘.. people need to know..’. ‘relevant’ in the third bullet point should be ‘relevance’.
Author’s feedback: Thank you, this has now been changed.

Results
1. Page 12, line 18 ‘party’ under face validity should be ‘partly’.

Author’s feedback: Thank you, this has now been changed.

2. Page 13, under ‘Preference of format and missing responses’ I think it would be good to have some examples of comments coming from the interviews with end users that support the summary given. This can either be a table or box with example comments, or perhaps a list of all of them in a supplementary document.
Author’s feedback: Thank you, as is mentioned in the methods section we did not transcribe the interviews as recording was not always possible. Furthermore, it is worth noting that these interviews are also more similar to user testing, in that they have a more technical focus and does not intend to explore people’s beliefs or attitudes. Instead, we sought to identify people’s understanding of formats by observing how they filled out the items, problems with terminology and people’s preferences. In these interviews the interaction between the interviewer and the user includes exploring the material at hand together, probing and observation. Potential problems or other issues were noted by the investigator in each setting and reported back to the working group, which considered the formats and terminology for revision. Many of these were item specific and included rewriting of a certain sentence for example.

3. Page 13: I would re-order the last sentence of the second paragraph to something like ‘Figure 2 shows the design changes used to avoid these problems’. Also, are the changes shown in figure 2 all the changes that were made to handle missing data, or were there others?
Page 14, first paragraph. ‘figure 3’ should be ‘figure 4’.
Authors feedback: Thank you, this has now been changed, the final changes are reported in figure 2.

Discussion
1. The background section says that the authors set out to develop the Claim Evaluation Tools to serve as the primary outcome measures to be used in future trials. The discussion section does not make explicit reference to how suitable or otherwise the authors now think the Tools are for this purpose. It would be useful for the authors to add a sentence or two that makes explicit reference to this aim.
Author’s feedback: Thank you, the discussion has now been revised to accommodate this

2. As mentioned earlier, I think the authors need to say something about the 10 items not tested (or make the text clearer so that my misunderstanding is not repeated by other readers).
Author’s feedback: this has now been explained in more detail in the discussion.

3. Page 14, line 26: there is an ‘a’ missing in ‘.. We believe that they will be A useful tool..’
Author’s feedback: The sentence has been deleted as part of other revisions.

4. Page 15: ‘.. measuring literacy skills..’ - do the authors have any thoughts on what level of literacy is required to successfully use the Tools in future evaluations? At present the authors say it would be important to measure literacy skills but do not say what level of minimum literacy is required to successfully use the Tools.
Author’s feedback: The items were tested in a low-income population with children from 10 years and up with English as their second language. However, we did not perform any formal literacy tests- we only included a small set of items to be included in the Rasch analysis to check for differential item functioning. Items that showed signs of this were deleted (described in the second paper).

Table 1
1. I think the item headings (e.g. 1. recognising the need for fair comparisons of treatments’) should be bolded, in italics or some similar formatting change so that it is clear that these are headings not items themselves.
Author’s feedback: Item headings are now in bold

2. I think the caption should be explicit fact 22/32 were tested here (or correct my misunderstanding).
Author’s feedback: See comment above regarding the items addressed in the pilots.